7. Methods for strong lens modelling

CHARLES KEETON

This chapter discusses computational and statistical methods for fitting models to strong lens data. It centres on parametric models of point-like lenses but includes extensions to composite models, free-form models, and extended sources. It describes how to use statistical tools including Monte Carlo Markov chains and nested sampling to explore the range of models that are consistent with data.

7.1. Introduction

Strong lensing is a versatile tool for astrophysics that can be used to study the physical properties and environments of lensing galaxies, to dissect the structure of source quasars and galaxies, to constrain cosmological parameters, and much more. Other chapters in this volume review the theory of strong lensing, the status of observations, and the variety of astrophysical applications that result. The goal of this chapter is to outline methods for fitting models to strong lens data. Since modelling is required for most applications of strong lensing, understanding the strengths and weaknesses of the analysis is key for drawing robust conclusions.

When discussing methodology, we need to distinguish between point-like and extended images. Point-like images (in a lensed quasar, for example) provide constraints on the potential and its derivatives at discrete positions, which can be described with a modest number of constraint equations or a straightforward χ^2 goodness of fit statistic. Established statistical methods can then be used to find the best fit and explore the range of allowed models. In this case the barrier to entry is low in the sense that fitting basic models does not require tremendous expertise, yet the potential for growth is high in the sense that advanced analyses can combine lensing with other astrophysical probes to draw conclusions that have broad reach. Extended images, by contrast, provide many more pixels of data but require many more free parameters (associated with the unknown shape of the source). Specialized methods must be used to simultaneously fit a mass model for the lens and a light model for the source. For pedagogical purposes, I focus on analysis methods that are applicable to point-like sources but include an overview of methods for modelling extended images (Section 7.5.4).

We also need to distinguish between what are traditionally known as parametric and free-form mass models, although (as we will see) the more meaningful division is between over-constrained and under-constrained models. When the model is over-constrained, it is usually straightforward to find the best fit and assess whether the model provides a good representation of the data. When the model is under-constrained, by contrast, there may be a large family of solutions that give perfect fits but have different levels of physical plausibility.

Parametric models can be simple (which may be 'good enough' for some applications), but they can also be sophisticated if they incorporate multiple mass components that are calibrated by astrophysical knowledge. Even so, the models may not capture all of the complexity of real galaxies, so it can be valuable to consider a much broader range of

Astrophysical Applications of Gravitational Lensing, ed. E. Mediavilla et al. Published by Cambridge University Press.

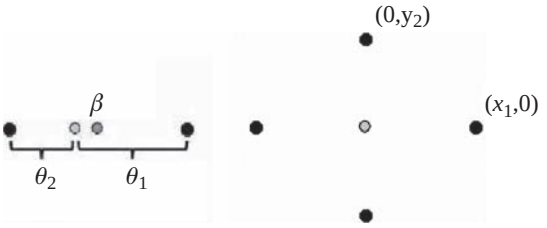

FIGURE 7.1. Left: simple example of a two-image lens. The grey circle indicates the lens, while the black circles indicate the images. The dark grey circle indicates the location of the background source. Right: example of a four-image lens in a cross configuration. Again the grey circle indicates the lens and the black circles indicate the images. The source is directly behind the lens.

possibilities through free-form models. I discuss basic parametric models in Section 7.3 and turn to more advanced models in Section 7.5.

Strong lens modelling has two stages that differ in how we approach the lens equation. In the 'forward problem', we postulate some fixed model (lens and source) and solve the lens equation to predict the images. In the 'inverse problem', we take the image data as known and reinterpret the lens equation in terms of constraints on model parameters. The two stages are often linked in practice, through least squares fitting. In this technique, we adopt some set of model parameters, use the forward problem to predict the images, compare the predictions to the data and then repeat the process until we find parameter values that minimize the differences. Since the forward and inverse stages both play a role in this analysis, it behoves us to understand both aspects of lens modelling.

To that end, we begin in Section 7.2 with some simple examples that can be solved analytically to illustrate the forward and inverse problems. In Section 7.3 we examine standard methods for handling more realistic scenarios, focusing on lenses with point-like images and parametric mass models. Then in Section 7.4 we take an extended excursion through statistics. No matter how good the data are, lens models always have some uncertainties, and much of our effort goes into characterizing and controlling them using statistical methods that are increasingly widespread not only in lensing but also throughout astrophysics. Finally, in Section 7.5 we return to strong lensing to discuss some advanced modelling techniques through a few case studies. Portions of this chapter draw on my review article (Keeton 2010), but on the whole this chapter is designed to be distinct and complementary.

7.2. Simple examples

Toy models are simplistic but nonetheless useful for illustrating concepts that we will see in more detail later. First consider a spherical lens that produces two images at angles θ_1 and θ_2 from the lens, as shown in Figure 7.1. For the moment, let us adopt a convention in which the angles $\theta_{1,2}$ are positive, as is the angle β between the lens and the source.

If the lens is a point mass with Einstein radius $\theta_{\rm E}$, the lens equation is

$$\beta = \theta - \frac{\theta_{\rm E}^2}{\theta}.$$

For the forward problem, we assume a value for β and then have a quadratic equation for θ; the two solutions correspond to the two image positions. Alternatively, for the inverse problem, we plug in the two image positions as $\theta = \theta_1$ and $\theta = -\theta_2$ (per our sign

convention) and then obtain two equations for the two unknowns $\theta_{\rm E}$ and β:

$$\beta = \theta_1 - \frac{\theta_{\rm E}^2}{\theta_1},$$
$$-\beta = \theta_2 - \frac{\theta_{\rm E}^2}{\theta_2}.$$

Adding the equations to eliminate β lets us solve for $\theta_{\rm E}$:

$$\theta_1 + \theta_2 = \theta_{\rm E}^2 \left(\frac{1}{\theta_1} + \frac{1}{\theta_2} \right) \quad \Rightarrow \quad \theta_{\rm E} = (\theta_1 \theta_2)^{1/2}. \tag{7.1}$$

In other words, for a point mass model the inferred Einstein radius is the geometric mean of the two image positions.

If the lens is instead a singular isothermal sphere (SIS), the lens equation is

$$\beta = \theta - \theta_{\rm E} \qquad (\theta > 0),$$
$$\beta = \theta + \theta_{\rm E} \qquad (\theta < 0).$$

The inverse problem has

$$\beta = \theta_1 - \theta_{\rm E},$$
$$-\beta = \theta_2 - \theta_{\rm E},$$

so the recovered Einstein radius is

$$\theta_{\rm E} = \frac{\theta_1 + \theta_2}{2}. \tag{7.2}$$

Thus, for an SIS model the inferred Einstein radius is the arithmetic mean of the two image positions (or half of the image separation).

Equations (7.1) and (7.2) show that we can get *different* answers from the *same* data depending on what we assume about the mass distribution (in this case, whether the mass is concentrated into a point mass or extended into an isothermal sphere). In practice, however, the model dependence is not especially strong for the Einstein radius. Suppose the lens is fairly symmetric, so the two image positions can be written as $\theta_1 = \theta_0 + \delta$ and $\theta_2 = \theta_0 - \delta$ such that δ is small. Then we can make a Taylor series expansion in δ to find:

$$\text{point mass:} \quad \theta_{\rm E} = (\theta_1 \theta_2)^{1/2} \approx \theta_0 \left[1 - \frac{\delta^2}{2\theta_0^2} + \mathcal{O}\left(\frac{\delta}{\theta_0} \right)^4 \right],$$
$$\text{isothermal:} \quad \theta_{\rm E} = \frac{\theta_1 + \theta_2}{2} \quad = \theta_0.$$

To a good approximation, the recovered Einstein radius is just the average of the image positions; the point mass model has a correction term, but that term is small. The Einstein radius, in other words, is not very sensitive to the choice of model.

Now consider an SIS lens with external tidal shear from mass nearby. We need to work with the full two-dimensional lens equation. Let $\mathbf{x}$ be a vector that denotes angular positions in the image plane (e.g. offsets in right ascension and declination), and $\mathbf{u}$ be an analogous vector in the source plane. In coordinates aligned with the shear, the lens equation has the form

$$\mathbf{u} = \mathbf{x} - \theta_{\rm E} \frac{\mathbf{x}}{|\mathbf{x}|} - \begin{bmatrix} \gamma & 0 \\ 0 & -\gamma \end{bmatrix} \mathbf{x},$$

where γ is the dimensionless strength of the shear. If the source is directly behind the lens ($\mathbf{u} = 0$), by symmetry there are two images on each axis. For the images on the x-axis, the y-component of the lens equation is satisfied trivially, and the non-trivial component is

$$0 = (1 - \gamma)x \mp \theta_{\rm E}.$$

(The $-$ sign applies when $x > 0$, while the $+$ sign applies when $x < 0$.) Similarly, for the images on the y-axis we have

$$0 = (1 + \gamma)y \mp \theta_{\rm E}.$$

The resulting cross configuration is shown in Figure 7.1. For the inverse problem, the only non-trivial equations are those for x_1 and y_2:

$$\theta_{\rm E} + \gamma x_1 = x_1,$$
$$\theta_{\rm E} - \gamma y_2 = y_2.$$

This is a system of equations for $\theta_{\rm E}$ and γ that can be represented by the matrix equation

$$\begin{bmatrix} 1 & x_1 \\ 1 & -y_2 \end{bmatrix} \begin{bmatrix} \theta_{\rm E} \\ \gamma \end{bmatrix} = \begin{bmatrix} x_1 \\ y_2 \end{bmatrix}$$

whose solution is

$$\theta_{\rm E} = \frac{2x_1 y_2}{x_1 + y_2} \quad \text{and} \quad \gamma = \frac{x_1 - y_2}{x_1 + y_2}.$$

The Einstein radius is a sort of average, while the shear is related to the asymmetry of the system.

If we model the same lens as a point mass with external shear, a similar analysis yields

$$\theta_{\rm E} = \left(\frac{2}{x_1^2 + y_2^2} \right)^{1/2} x_1 y_2 \quad \text{and} \quad \gamma = \frac{x_1^2 - y_2^2}{x_1^2 + y_2^2}.$$

To understand the differences between models, suppose again that the cross is fairly symmetric so we can write $x_1 = \theta_0 + \delta$ and $y_2 = \theta_0 - \delta$ with δ small. Then we can use a Taylor series expansion to find for the Einstein radius:

$$\text{point mass:} \quad \theta_{\rm E} \approx \theta_0 \left[1 - \frac{3\delta^2}{2\theta_0^2} + \mathcal{O}\left(\frac{\delta}{\theta_0} \right)^4 \right],$$
$$\text{isothermal:} \quad \theta_{\rm E} \approx \theta_0 \left[1 - \frac{\delta^2}{\theta_0^2} + \mathcal{O}\left(\frac{\delta}{\theta_0} \right)^4 \right].$$

Similarly, we find for the shear:

$$\text{point mass:} \quad \gamma \approx \frac{2\delta}{\theta_0} + \mathcal{O}\left(\frac{\delta}{\theta_0} \right)^3,$$
$$\text{isothermal:} \quad \gamma \approx \frac{\delta}{\theta_0} + \mathcal{O}\left(\frac{\delta}{\theta_0} \right)^3.$$

There are differences between the results from the two models, but for $\theta_{\rm E}$ they are only in the correction term, while for γ they are in the leading term. In other words, the Einstein radius is fairly robust, but the shear is more sensitive to assumptions built into the lens model.

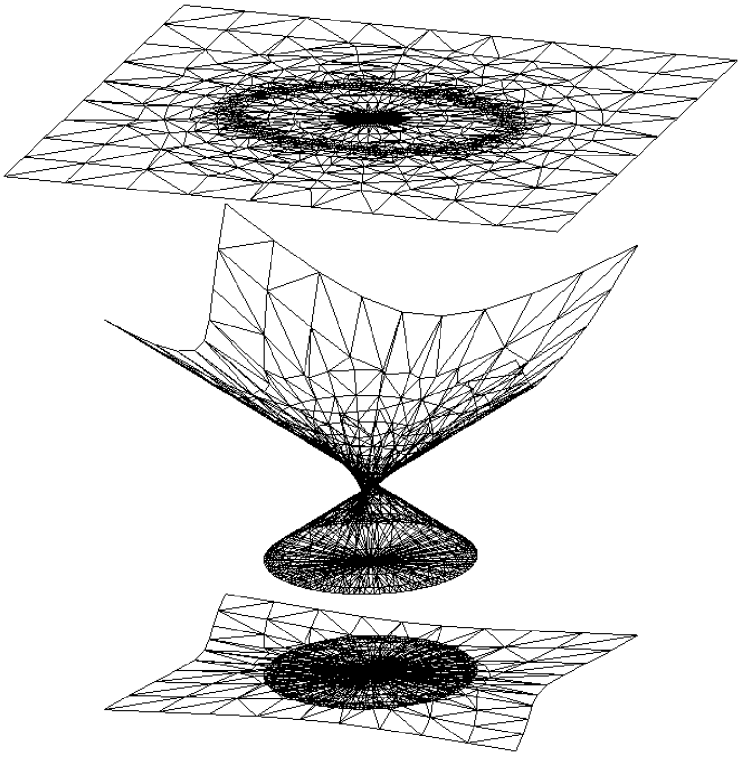

FIGURE 7.2. Depiction of the mapping from tiles in the image plane (top) to tiles in the source plane (bottom). In the middle image, height has been added to illustrate how the lens mapping effectively takes the image plane and folds it over on itself to create the multiply imaged region.

7.3. Basic methods

In the preceding examples we were able to solve the equations analytically for both the forward and inverse problems. In most real lenses, however, that is not the case, so we need some computational algorithms and statistical techniques to find images and fit models.

7.3.1 *Forward problem*

Often the lens equation is a non-linear function of $\mathbf{x}$ that can only be solved numerically. We can build a general algorithm for finding images by covering the image plane with tiles and then using the lens equation to map each tile back to the source plane. Multiply imaged regions in the source plane will be covered by overlapping tiles, as shown in Figure 7.2. To solve the lens equation, we identify tiles that overlie the source position and run a numerical root finder in the corresponding image plane tiles to obtain the exact image positions. For practical reasons we may want to construct tiles using polar coordinates near the centres of lens galaxies (to resolve their features) but Cartesian coordinates at large radii (where polar tiles would become large in the azimuthal direction). We can use adaptive subgridding near critical curves to obtain good resolution in regions where the distortion and folding of the lens mapping are the most extreme. These steps may yield a rather complicated distribution of points in the image plane (see Figure 7.3),

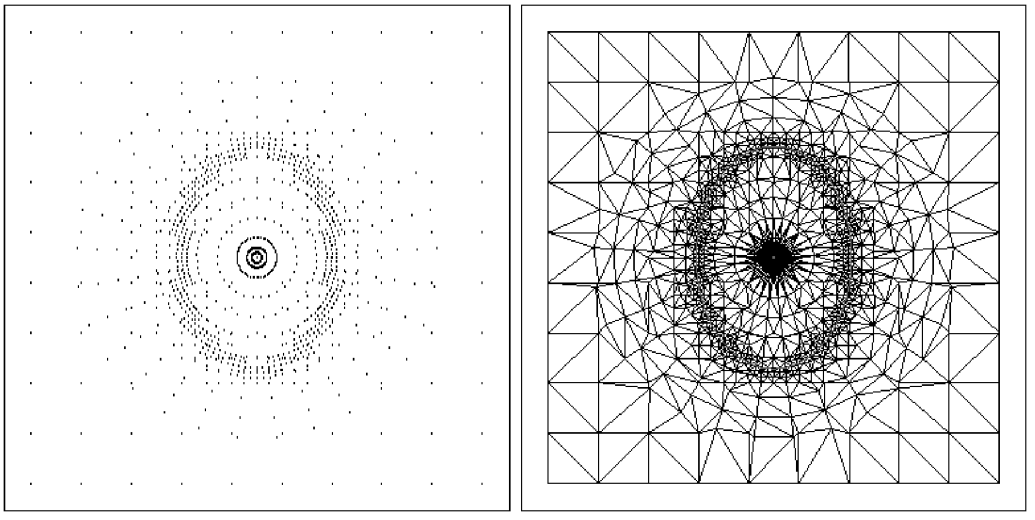

FIGURE 7.3. Illustration of the tiling using to solve the lens equation. The left panel shows a set of points obtained by combining a polar grid centred on the lens galaxy with a background Cartesian grid and using adaptive subgridding near critical curves. The right panel shows a Delaunay triangulation (Shewchuk 1996, 2002) of the grid points.

but the points can be connected into a coherent tiling using Delaunay triangulation (Shewchuk 1996, 2002). (See Keeton 2010 for more discussion of these points.)

Once we find the positions of images, we can predict the magnifications and time delays using standard results from lens theory. Distortions are described by the magnification tensor,

$$\boldsymbol{\mu} = \frac{\partial \mathbf{x}}{\partial \mathbf{u}} = \left(\frac{\partial \mathbf{u}}{\partial \mathbf{x}} \right)^{-1} = \begin{bmatrix} 1 - \phi_{xx} & -\phi_{xy} \\ -\phi_{xy} & 1 - \phi_{yy} \end{bmatrix}^{-1} , \tag{7.3}$$

where subscripts denote partial derivatives (e.g. $\phi_{xx} = \partial^2 \phi / \partial x^2$). The scalar magnification is then

$$\mu = \det \boldsymbol{\mu} = \left[(1 - \phi_{xx})(1 - \phi_{yy}) - \phi_{xy}^2 \right]^{-1} . \tag{7.4}$$

As written, μ is a signed quantity such that the sign indicates the parity of the image. The general expression for the excess light travel time (relative to an unlensed ray) is

$$t(\mathbf{x}; \mathbf{u}) = t_0 \left[\frac{1}{2} |\mathbf{x} - \mathbf{u}|^2 - \phi(\mathbf{x}) \right] \quad \text{where} \quad t_0 = \frac{1 + z_\mathrm{L}}{c} \frac{D_\mathrm{L} D_\mathrm{S}}{D_\mathrm{LS}} . \tag{7.5}$$

Here D_L and D_S are angular diameter distances to the lens and source, while D_LS is the angular diameter distance from the lens to the source. The differential time delay between images at positions $\mathbf{x}_i$ and $\mathbf{x}_j$ is then

$$\Delta t_{ij} = t_0 \left[\frac{|\mathbf{x}_i|^2 - |\mathbf{x}_j|^2}{2} - (\mathbf{x}_i - \mathbf{x}_j) \cdot \mathbf{u} - \phi(\mathbf{x}_i) + \phi(\mathbf{x}_j) \right] . \tag{7.6}$$

7.3.2 *Inverse problem*

In the inverse problem, it may be impossible to find *any* solution (either analytic or numerical) that solves all of the constraint equations simultaneously. There may be more constraints than free parameters that can be adjusted to seek a solution. The model

may simply be incapable of reproducing the data (i.e. fundamentally wrong). Even if the model is correct, the predicted and observed data may not match because of noise in the measurements. We therefore avoid trying to solve the constraint equations exactly and instead try to minimize the difference between the model and data. In 'least squares fitting', we define a goodness of fit that has the form

$$\chi^2 = \sum \frac{(\text{model} - \text{data})^2}{(\text{uncertainties})^2}, \tag{7.7}$$

where the sum runs over all of the observables. We can then do two things: we can look for the best fit (corresponding to the minimum value of χ^2), and we can explore the range of 'allowed' models (corresponding to the region where the χ^2 value is acceptable). In Section 7.4 we discuss the statistical framework for exploring the parameter space. Here we specify how strong lens data provide constraints on models. The main data include the image positions, fluxes and time delays. The list of free parameters includes not only those associated with the mass model (see Section 7.3.3), but also the position and flux of the source. (This presentation draws on the review by Keeton 2010.)

Positions. We impose constraints from image positions with a χ^2 term of the form

$$\chi^2_{\rm pos} = \sum_{\rm i} \left(\mathbf{x}_i^{\rm mod} - \mathbf{x}_i^{\rm obs}\right)^t \, \mathbf{S}_i^{-1} \, \left(\mathbf{x}_i^{\rm mod} - \mathbf{x}_i^{\rm obs}\right), \tag{7.8}$$

where the sum runs over the images. We characterize the uncertainties using the covariance matrix $\mathbf{S}_i$ to allow for the possibility that the astrometric error bars involve an error ellipse. If the error bar is σ_i in both x and y directions, the covariance matrix is

$$\mathbf{S}_i = \begin{bmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_i^2 \end{bmatrix}$$

and $\chi^2_{\rm pos}$ has the form of equation (7.6). We can obtain an alternative expression for $\chi^2_{\rm pos}$ as follows (see Kochanek 1991). Let us define the source position associated with each observed image,

$$\mathbf{u}_i^{\rm obs} = \mathbf{x}_i^{\rm obs} - \boldsymbol{\alpha}\left(\mathbf{x}_i^{\rm obs}\right).$$

The predicted image positions are related to the model source position by

$$\mathbf{u}^{\rm mod} = \mathbf{x}^{\rm mod} - \boldsymbol{\alpha}\left(\mathbf{x}^{\rm mod}\right).$$

If we subtract these two equations, make a Taylor series expansion of $\boldsymbol{\alpha}(\mathbf{x})$ and use equation (7.3) to simplify, we obtain

$$\delta\mathbf{u}_i = \delta\mathbf{x}_i - \left[\boldsymbol{\alpha}\left(\mathbf{x}_i^{\rm mod}\right) - \boldsymbol{\alpha}\left(\mathbf{x}_i^{\rm obs}\right)\right] \approx \boldsymbol{\mu}_i^{-1}\,\delta\mathbf{x}_i,$$

where $\delta\mathbf{x}_i = \mathbf{x}_i^{\rm mod} - \mathbf{x}_i^{\rm obs}$ and $\delta\mathbf{u}_i = \mathbf{u}^{\rm mod} - \mathbf{u}_i^{\rm obs}$. Thus, to lowest order we can replace each $\delta\mathbf{x}_i$ factor in equation (7.8) with $\boldsymbol{\mu}_i\,\delta\mathbf{u}_i$ and write

$$\chi^2_{\rm pos} \approx \sum_i \left(\mathbf{u}^{\rm mod} - \mathbf{u}_i^{\rm obs}\right)^t \, \boldsymbol{\mu}_i^t\,\mathbf{S}_i^{-1}\boldsymbol{\mu}_i \, \left(\mathbf{u}^{\rm mod} - \mathbf{u}_i^{\rm obs}\right). \tag{7.9}$$

Which expression for $\chi^2_{\rm pos}$ should we use? Equation (7.8) has the virtue of being exact and ensuring that the predicted number of images matches the observed number. The main downside is that it requires solving the lens equation, which involves the computational effort of tiling the image and source planes. Equation (7.9) needs only a handful of deflection calculations to map the observed image positions back to the source plane, so

it can be much faster. Moreover, equation (7.9) is quadratic in $\mathbf{u}^{\rm mod}$ so we can find the best fit value analytically by solving $\nabla_{\mathbf{u}}\chi^2_{\rm pos} = 0$, which is equivalent to

$$\mathbf{A}\,\mathbf{u}^{\rm mod} = \mathbf{b}, \tag{7.10}$$

where

$$\mathbf{A} = \sum_i \boldsymbol{\mu}_i^t\,\mathbf{S}_i^{-1}\,\boldsymbol{\mu}_i \quad \text{and} \quad \mathbf{b} = \sum_i \boldsymbol{\mu}_i^t\,\mathbf{S}_i^{-1}\,\boldsymbol{\mu}_i\,\mathbf{u}_i^{\rm obs}.$$

Solving the matrix equation (7.10) is much faster than explicitly searching the $\mathbf{u}^{\rm mod}$ parameter space. The approximation that underlies equation (7.9) becomes increasingly accurate as the residuals decrease, so the expression is reliable near the minimum of the χ^2 surface (which is the region of interest).

There is, however, one important concern with equation (7.9): it does not explicitly check that the model predicts the correct number of images. There is an implicit penalty for models that predict fewer images than observed: in such a case, some of the $\mathbf{u}_i^{\rm obs}$ values should wind up far from the others, leading to a large $\chi^2_{\rm pos}$ value that would indicate a bad model. More troublesome is the possibility that the model could predict more images than observed and we would have no way of knowing based on the value of $\chi^2_{\rm pos}$ from equation (7.9). In my experience, though, there is little ambiguity about the number of images when the lensing is dominated by a single galaxy; it is difficult if not impossible to construct a four-image model in which two of the images closely match an observed two-image configuration while two of the images are spurious. Therefore equation (7.9) may be adequate in 'standard' cases of galaxy-scale strong lensing. The situation becomes more complicated if there are multiple galaxies that are close enough for the caustics to interact and create either unusual two- or four-image configurations or lenses with different numbers of images altogether. In such cases the safe approach is to work with equation (7.8) and accept the additional runtime for the sake of being confident that the model predicts the correct number of images.

Brightnesses. Constraints on the brightnesses (or fluxes) enter through the χ^2 term:

$$\chi^2_{\rm flux} = \sum_i \frac{\left(F_i^{\rm obs} - \mu_i F^{\rm src}\right)^2}{\sigma_{f,i}^2}. \tag{7.11}$$

The units are arbitrary because they appear in both the numerator and denominator and thus factor out. Therefore, brightness constraints can be imposed using either absolute fluxes or flux ratios between images. The optimal source flux is found by solving $d\chi^2_{\rm flux}/dF^{\rm src} = 0$:

$$F^{\rm src} = \frac{\sum_i F_i^{\rm obs}\,\mu_i/\sigma_{f,i}^2}{\sum_i \mu_i^2/\sigma_{f,i}^2}. \tag{7.12}$$

Time delays. To specify constraints on time delays, it is convenient to rewrite equation (7.5) as

$$t_i^{\rm mod} = t_0\tau_i^{\rm mod} + T_0, \tag{7.13}$$

where we explicitly include a time zero point, T_0, and we write the model prediction in two pieces: the (dimensionless) factors that depend on the lens model appear in

$$\tau_i^{\rm mod} = \frac{1}{2}\left|\mathbf{x}_i^{\rm mod} - \mathbf{u}^{\rm mod}\right|^2 - \phi\left(\mathbf{x}_i^{\rm mod}\right),$$

while the physical factors that depend on cosmology are

$$t_0 = \frac{1+z_\mathrm{L}}{c}\,\frac{D_\mathrm{L} D_\mathrm{S}}{D_\mathrm{LS}}.$$

Using equation (7.13), we can write the χ^2 term for time delay constraints as

$$\chi^2_\mathrm{tdel} = \sum_i \frac{\left(t_i^\mathrm{obs} - t_0 \tau_i^\mathrm{mod} - T_0\right)^2}{\sigma^2_{t,i}}. \tag{7.14}$$

This expression is quadratic in t_0 and T_0, so we can find the optimal values by simultaneously solving $\partial\chi^2_\mathrm{tdel}/\partial t_0 = 0$ and $\partial\chi^2_\mathrm{tdel}/\partial T_0 = 0$, which is equivalent to

$$\begin{bmatrix} \sum_i \frac{\left(\tau_i^\mathrm{mod}\right)^2}{\sigma^2_{t,i}} & \sum_i \frac{\tau_i^\mathrm{mod}}{\sigma^2_{t,i}} \\ \sum_i \frac{\tau_i^\mathrm{mod}}{\sigma^2_{t,i}} & \sum_i \frac{1}{\sigma^2_{t,i}} \end{bmatrix} \begin{bmatrix} t_0 \\ T_0 \end{bmatrix} = \begin{bmatrix} \sum_i \frac{\tau_i^\mathrm{mod} t_i^\mathrm{obs}}{\sigma^2_{t,i}} \\ \sum_i \frac{t_i^\mathrm{obs}}{\sigma^2_{t,i}} \end{bmatrix}. \tag{7.15}$$

7.3.3 *Parametric models*

Let us briefly consider what we want to constrain, focusing for now on basic parametric models. The simplest realistic mass model for a lens galaxy has elliptical symmetry[†] and a power law density profile that may or may not have a finite density core. In coordinates aligned with the major axis of the ellipse, the scaled surface mass density can be written as

$$\kappa(x,y) = \frac{\Sigma(x,y)}{\Sigma_\mathrm{crit}} = \frac{b^{2-\alpha}}{2\left(s^2 + x^2 + y^2/q^2\right)^{1-\alpha/2}}, \tag{7.16}$$

where b is a normalization factor defined to have dimensions of length, s is the core radius, $q \leq 1$ is the axis ratio of the ellipse, and α is the power law index defined such that asymptotically the enclosed mass scales as

$$M(r) \propto r^\alpha \quad \text{where} \quad \alpha \begin{cases} < 1 & \text{steeper than isothermal,} \\ = 1 & \text{isothermal,} \\ > 1 & \text{shallower than isothermal.} \end{cases}$$

A number of other parametric models have been used over the years; Keeton (2001) gives a catalogue of many common choices. Multiple parametric components can be combined to build composite models that become arbitrarily complex (see Section 7.5.1).

Mass near the main lens galaxy or along the line of sight can affect the images at levels larger than the noise. For perturbers that lie in the same plane as the main galaxy[‡] and are 'far' from the lens (compared with the Einstein radius), we can make a Taylor series expansion of the lens potential from the environment:

$$\phi_\mathrm{env} = \frac{\kappa_c}{2} r^2 + \frac{\gamma}{2} r^2 \cos 2(\theta - \theta_\gamma) + \cdots. \tag{7.17}$$

The zeroth and first order terms are irrelevant because they represent the zero point of the potential and a translation of the source plane, respectively. The second order κ_c term corresponds to a uniform mass sheet; this term must be treated with care because any mass sheet can be absorbed by rescaling the mass of the lens, the flux of the source,

[†] In some cases it may be adequate to use a circular mass distribution along with external shear, but elliptical mass distributions are more generic, and both ellipticity and shear are required in many real lenses.

[‡] See Section 7.5.5 for remarks about handling effects from mass along the line of sight.

and the time delays (Falco, Gorenstein and Shapiro 1985; Gorenstein, Shapiro and Falco 1988). The γ term corresponds to an external tidal shear, which is detectable and often required to obtain good fits. Terms at third order and higher are generally small unless the neighbours are close to the lens, but given the quality of modern data they may not be negligible (see Section 7.5.1). Nevertheless, for a minimal model it is common to include only the shear term.

Before going further it is instructive to pause and count constraints and free parameters. For a lens with two or four point-like images, we can usually measure the position of the lens galaxy ($\mathbf{x}_{\rm gal}$) along with the positions and fluxes of the images ($\mathbf{x}_i$ and F_i). We may or may not be able to measure all of the differential time delays (Δt_i). The constraints we could obtain for 'standard' double and quad lenses are therefore:

	$\mathbf{x}_{\rm gal}$	$\mathbf{x}_i$	F_i	Δt_i	total
double	2	2×2	2	1	9
quad	2	4×2	4	3	17

The list of unknown parameters includes the source position and flux ($\mathbf{u}_{\rm src}$ and $F_{\rm src}$) and the position of the main lens galaxy ($\mathbf{x}_{\rm gal}$). A parametric mass model has at least three free parameters ($\mathbf{q}_{\rm gal}$): a normalization parameter, an ellipticity and a position angle. The environment has at least two free parameters ($\mathbf{q}_{\rm env}$): the strength and direction of the tidal shear. More complex models can have even more unknown parameters. If we have time delay constraints, we also need to fit for the time scale t_0. The free parameters are therefore:

$\mathbf{u}_{\rm src}$	$F_{\rm src}$	$\mathbf{x}_{\rm gal}$	$\mathbf{q}_{\rm gal}$	$\mathbf{q}_{\rm env}$	t_0	total
2	1	2	≥ 3	≥ 2	1	≥ 11

Doubles are under-constrained if we include both ellipticity and shear in the model; we will need some independent constraints or priors in order to obtain useful results. By contrast, quads are over-constrained so we can expect to find well-defined best fit models and determine whether the models are objectively acceptable.

7.4. Statistical framework

We now turn to statistical methods for fitting models and constraining parameters. The techniques discussed here are not unique to lensing, but they are becoming increasingly prevalent in strong lens modelling so it is worthwhile to present them as a unit.

7.4.1 *Likelihood*

All measurements have some amount of uncertainty that must be factored into the analysis. If the measurement noise follows a Gaussian distribution with standard deviation σ (often a reasonable assumption), we can write down the probability distribution for the measured value ($d^{\rm obs}$) of some observable, given the model prediction ($d^{\rm mod}$):

$$\mathcal{L}\left(d^{\rm obs}|d^{\rm mod}\right) \propto \exp\left[-\frac{\left(d^{\rm obs}-d^{\rm mod}\right)^2}{2\sigma^2}\right]. \tag{7.18}$$

This is called the 'likelihood of the data given the model' and it is depicted in Figure 7.4a. Notice that the exponent has the same form as the general χ^2 that we wrote down in equation (7.7), so for Gaussian noise we have a relation between likelihood and χ^2:

$$\mathcal{L} \propto {\rm e}^{-\chi^2/2} \quad \Leftrightarrow \quad \chi^2 = -2\ln\mathcal{L} + {\rm const.} \tag{7.19}$$

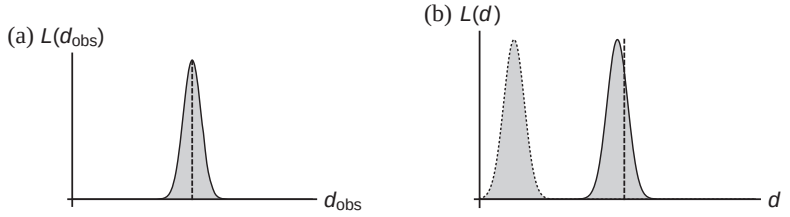

FIGURE 7.4. (a) The observed value $d_{\rm obs}$ may not exactly match model predictions (dashed line) because of measurement noise. (b) If a model has a likelihood distribution far from the observed value (dashed line), the model is a poor fit to the data. We try to adjust the model until the likelihood distribution is closer to the observed value.

We can use χ^2 and $\mathcal{L}$ somewhat interchangeably as long as we remember that we want χ^2 to be *low* and $\mathcal{L}$ to be *high*.

If the model predictions depend on some parameters q,[†] we can write the likelihood as

$$\mathcal{L}\left(d^{\rm obs}|q\right) \propto \exp\left[-\frac{\left(d^{\rm obs}-d^{\rm mod}(q)\right)^2}{2\sigma^2}\right].$$

Now we have a way to look for the best values of the parameters: when the fit is poor, the predictions $d^{\rm mod}$ are far from the observed values $d^{\rm obs}$, so χ^2 is high and the likelihood is low (see the left distribution in Figure 7.4b). We try to adjust the model parameters to reduce χ^2 and increase the likelihood (see the distribution in Figure 7.4b). This is known as the 'maximum likelihood method' for model fitting.

We can always find a *best* fit, but we still need to ask whether it is a *good* fit. In particular, we need to decide whether differences between the model and data are small enough to be consistent with measurement noise or large enough to indicate that the model does not provide an adequate description of the data. The first step is to quantify the number of 'degrees of freedom',

$$\nu = (\#\text{ constraints}) - (\#\text{ free parameters}). \tag{7.20}$$

If the noise is Gaussian then χ^2 will follow a characteristic probability distribution that depends on ν (e.g. Press et al. 1992):

$$P(\chi^2|\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\,(\chi^2)^{\nu/2-1}\mathrm{e}^{-\chi^2/2}. \tag{7.21}$$

This probability distribution is shown in Figure 7.5 for a few different values of ν. The distribution has mean

$$\langle\chi^2\rangle = \nu$$

and the peak is located at

$$\chi^2_{\rm peak} = \max(\nu-2, 0).$$

Roughly speaking, then, χ^2 should be comparable to the number of degrees of freedom for a 'good' fit; conversely, a value $\chi^2 \gg \nu$ indicates a 'bad' fit. This is just a rule of thumb, though. Given statistical scatter in the noise, χ^2 can be a little larger or smaller than ν even for a good fit. Strictly speaking, we must use the full χ^2 probability distribution

[†] There may be multiple parameters that can be collected into a parameter vector, but we omit vector notation here for simplicity.

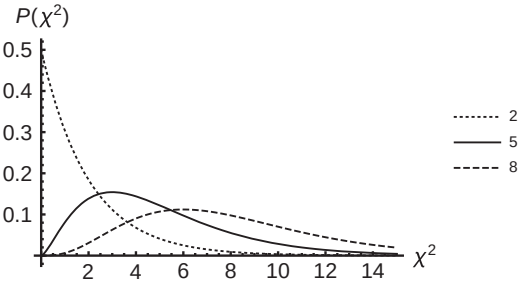

FIGURE 7.5. χ^2 probability distribution for $\nu = 2$ (dotted), 5 (solid), and 8 (dashed).

(7.21) to evaluate whether a particular χ^2 value does or does not qualify as an acceptable fit. For example, if the tail of the distribution contains less than 5% of the probability, then we can say a model is ruled out at the 95% confidence level.

7.4.2 *Covariance*

Often the measurement uncertainties in different quantities are statistically independent, but sometimes they are correlated. (An example is error ellipses in astrometry, which could arise from the shape of the beam in radio astronomy.) To quantify a relation between two quantities x and y, we use the statistical covariance:

$$\begin{aligned} \mathrm{Cov}(x,y) &= \langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle \\ &= \langle xy - \langle x \rangle y - x \langle y \rangle + \langle x \rangle \langle y \rangle \rangle \\ &= \langle xy \rangle - \langle x \rangle \langle y \rangle . \end{aligned}$$

The usual variance is $\sigma_x^2 = \mathrm{Cov}(x,x)$. If we have an array of data $\mathbf{d} = (d_1, d_2, d_3, \ldots)$, we define the covariance matrix

$$\mathbf{C} = \begin{bmatrix} \sigma_1^2 & \mathrm{Cov}(d_1,d_2) & \mathrm{Cov}(d_1,d_3) & \cdots \\ \mathrm{Cov}(d_2,d_1) & \sigma_2^2 & \mathrm{Cov}(d_2,d_3) & \\ \mathrm{Cov}(d_3,d_1) & \mathrm{Cov}(d_3,d_2) & \sigma_3^2 & \\ \vdots & & & \ddots \end{bmatrix} .$$

We can then define a generalized goodness of fit that handles correlated data:

$$\chi^2 = \left(\mathbf{d}^{\mathrm{obs}} - \mathbf{d}^{\mathrm{mod}}\right)^t \mathbf{C}^{-1} \left(\mathbf{d}^{\mathrm{obs}} - \mathbf{d}^{\mathrm{mod}}\right) . \tag{7.22}$$

Everything we saw in Section 7.4.1 about the χ^2 distribution still applies. If the data are independent, $\mathbf{C}$ is diagonal and equation (7.22) reduces to

$$\chi^2 = \sum_i \frac{\left(d_i^{\mathrm{obs}} - d_i^{\mathrm{mod}}\right)^2}{\sigma_i^2} . \tag{7.23}$$

7.4.3 *Optimizing parameters*

Once we define the model and its likelihood function, our first task is to find the parameters that yield the best fit. This amounts to finding the peak in the likelihood function, or equivalently the minimum in the χ^2 function. For this analysis there is an

important distinction between parameters that have a linear dependence in the model and parameters that are non-linear.

Linear parameters. Consider a situation in which x is an independent variable and the model prediction is a straight line of the form

$$d^{\rm mod} = mx + b$$

for some parameters m and b. Then χ^2 has the form

$$\chi^2 = \sum_i \frac{\left(d_i^{\rm obs} - mx_i - b\right)^2}{\sigma_i^2}.$$

This function is quadratic in both m and b, so we can find the minimum by solving

$$0 = \frac{\partial \chi^2}{\partial m} = -2\sum_i \frac{x_i\left(d_i^{\rm obs} - mx_i - b\right)}{\sigma_i^2},$$

$$0 = \frac{\partial \chi^2}{\partial b} = -2\sum_i \frac{\left(d_i^{\rm obs} - mx_i - b\right)}{\sigma_i^2}.$$

While this might look a little complicated, it is just a pair of linear equations with two unknowns m and b. It can be written as a matrix equation,

$$\begin{bmatrix} \sum_i \frac{x_i^2}{\sigma_i^2} & \sum_i \frac{x_i}{\sigma_i^2} \\ \sum_i \frac{x_i}{\sigma_i^2} & \sum_i \frac{1}{\sigma_i^2} \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} \sum_i \frac{x_i d_i^{\rm obs}}{\sigma_i^2} \\ \sum_i \frac{d_i^{\rm obs}}{\sigma_i^2} \end{bmatrix}$$

and then solved by standard matrix inversion. This approach can be applied to arbitrary sets of linear parameters; we have already seen examples of it in Section 7.3.2 with $\mathbf{u}^{\rm mod}$, $F^{\rm src}$, t_0 and T_0.

Non-linear parameters. Unfortunately, there is no comparable analysis that automatically reveals the optimal values of non-linear parameters. We must explicitly search the parameter space to look for the peak in the likelihood function or the minimum in the χ^2 function. Algorithms to search for the minimum of a function in multiple dimensions are well established (e.g. Press et al. 1992), but they can face certain challenges. The χ^2 function may have several local minima in addition to the global minimum (see Figure 7.6a). Most optimization algorithms are susceptible to getting stuck in local minima, so it is important to understand the structure of the χ^2 surface that exists in each particular problem. If a multidimensional χ^2 surface has long, narrow valleys (e.g. Figure 7.6b), the optimization algorithm must either take small steps (and thus become inefficient) or figure out how to use different step sizes in different directions. The challenge is even more acute if the valley is curved (e.g. Figure 7.6c).

Combinations of linear and non-linear parameters. It is worth noting that the two techniques can be combined when optimizing a model that has both linear and non-linear parameters. For example, suppose we have parameters a and b such that

$$d^{\rm mod} = a\, f(b)$$

and χ^2 has the form

$$\chi^2(a,b) = \sum \frac{\left[af(b) - d^{\rm obs}\right]^2}{\sigma^2}.$$

FIGURE 7.6. (a) A χ^2 function in 1-D that exhibits local minima. (b) A χ^2 function in 2D that exhibits a narrow, straight valley. (c) A χ^2 function in 2D that exhibits a long, curved valley.

We can solve analytically for the optimal value of a at a given value of b:

$$0 = \frac{\partial \chi^2}{\partial a} \quad \Rightarrow \quad a_{\rm opt}(b) = \frac{\sum f(b)\, d^{\rm obs}/\sigma^2}{\sum f(b)^2/\sigma^2}.$$

Then we can think of χ^2 as a function of just the non-linear parameter:

$$\chi^2(b) = \chi^2\left(a_{\rm opt}(b), b\right).$$

This reduces the number of parameters that we have to search explicitly. It is valid whenever we want to optimize the linear parameter(s). It will not, however, properly capture the uncertainties in a (see Section 7.4.6).

7.4.4 *Basic error bars*

After finding the best fit, our next task is to determine the range over which models are acceptable; this defines the error bars on the inferred parameter values. To specify what we mean by 'acceptable' let us consider the simple case of a 1D Gaussian distribution:

$$\mathcal{L} \propto {\rm e}^{-\chi^2/2} \quad \text{where} \quad \chi^2 = \frac{(x-d)^2}{\sigma^2}.$$

We conventionally quote the error bars in terms of σ, but how can we think about them more generally? If we vary x by $\pm 1\sigma$, that corresponds to changing χ^2 by $\Delta\chi^2 = 1$. If instead we vary x by $\pm 2\sigma$, that corresponds to $\Delta\chi^2 = 4$. We could therefore choose to define general error bars in terms of $\Delta\chi^2$ thresholds.

However, it is more meaningful to think about area under the curve, because integrating a probability distribution yields a probability that can be interpreted in familiar ways. With a Gaussian distribution, integrating over the region $\pm\sigma$ around the median encompasses 68% of the total probability (see Figure 7.7a). Thus, there is a 68% chance that the true value lies in this range, so we call this the 68% confidence interval. Similarly, the region within $\pm 2\sigma$ of the median is the 95% confidence interval.

This provides a better way to generalize the notion of error bars to distributions that may not be Gaussian. For a 1D distribution, we can quote the median value along with the 68% confidence interval defined such that the left and right tails each contain 16% of the probability. Figure 7.7b shows an example for a skewed distribution. Another common range to quote is the 95% confidence interval defined such that the left and right tails each contain 2.5% of the probability.

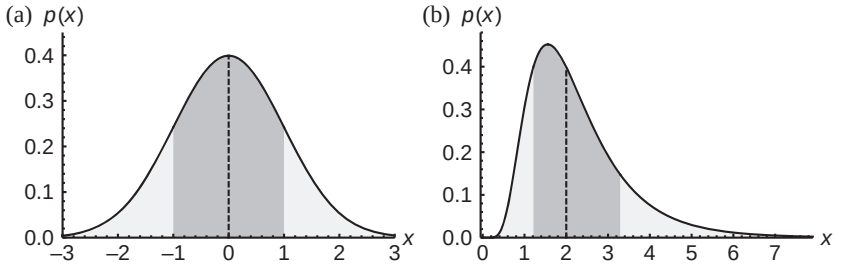

FIGURE 7.7. (a) With a Gaussian distribution, the region $\pm\sigma$ around the median contains 68% of the probability (dark grey), while the left and right tails each contain 16% of the probability (light grey). (b) Similarly, we can characterize a general 1-D distribution using the median value (dotted line) along with the 68% confidence interval (dark grey) defined such that the left and right tails each contain 16% of the probability (light grey).

7.4.5 *Bayesian statistics*

In thinking about error bars we have begun to talk about probabilities. If we take the plunge and fully adopt the language of probability, we can use Bayesian statistics as a powerful framework for constraining model parameters and even comparing different models in an objective way (Gelman et al. 2003).

Let us review some key aspects of probability theory (e.g. Ross 2012). If we have two random variables a and b, we can discuss three different probability distributions. The joint distribution $p(a, b)$ describes all possible values that a and b can have. The conditional distribution $p(a|b)$ describes the values that a can adopt if we specify that b has some given value. Finally, the marginal distribution $p(a)$ describes the values that a can adopt if we do not care at all about b (meaning that we let b take on any value it wants). The marginal distribution for a is obtained by integrating over b:

$$p(a) = \int p(a, b)\ db.$$

The joint, conditional and marginal distributions are related by

$$p(a, b) = p(a|b)\ p(b). \tag{7.24}$$

The various distributions are illustrated in Figure 7.8.

We can apply these ideas to modelling: the data d and model parameters q are formally considered random variables because of noise.† Using equation (7.24), we can relate the joint distribution $p(d, q)$ to either of the conditional distributions $p(d|q)$ or $p(q|d)$ as follows:

$$p(d, q) = p(q|d)\, p(d) = p(d|q)\, p(q).$$

This leads immediately to Bayes's theorem (e.g. Gelman et al. 2003):

$$p(q|d) = \frac{p(d|q)\, p(q)}{p(d)}. \tag{7.25}$$

As we have seen, this relation follows directly from probability theory. It becomes the foundation for Bayesian statistics when we give statistical interpretations to the various quantities:

† As before, d and q can be vector quantities.

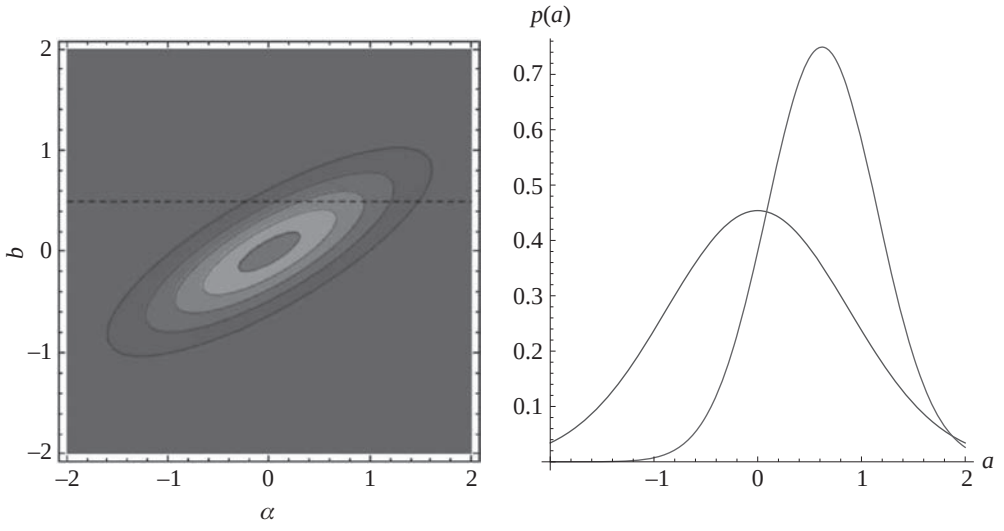

FIGURE 7.8. The left panel shows a sample joint probability distribution for two random variables a and b. In the right panel, the red curve shows the conditional distribution $p(a|b)$ given that b has the value indicated by the dotted line in the left panel, while the blue curve shows the marginal distribution $p(a)$.

- $p(d|q)$ is the probability of the data given the model; it is identical to what we have been calling the likelihood, so we re-label it $\mathcal{L}(d|q)$.
- $p(q)$ is the 'prior' probability distribution for q; it encodes any knowledge or assumptions that we have about q before applying the new constraints.
- $p(d)$ is called the 'evidence'; we will see more about this in a moment.
- $p(q|d)$ is the 'posterior' probability distribution for the model given the data.

This approach provides the framework for two types of analysis.

Inference. One thing we want to do is understand the allowed range of parameter values for a given model. The posterior provides a complete description of the probability distribution for the model parameters as constrained by the data. It does so without requiring any simplifying assumptions about the shape of the distribution (such as Gaussianity). The posterior naturally encodes any covariances or degeneracies between parameters and any complexities such as multimodality. The posterior, in other words, contains everything we want to know about the model parameters. (The challenge, as we will see, is figuring out how to access that information.)

Model comparison. A second type of analysis arises if we want to compare several distinct models to see which is better. Suppose we have two classes of models whose likelihood curves are shown in Figure 7.9. (Let the priors be uniform over the relevant range of parameters.) The peak likelihoods are the same, so the best fit models are equally good, but the blue likelihood curve is wider than the red one. Which model class is better?

If we were trying to constrain parameter values, we might prefer to see the shaded model because it provides tighter constraints. But if we are trying to figure out which model provides an overall better description of the data, we must recognize that the shaded model requires more fine tuning; the parameters have to be in a fairly narrow range in order to provide a reasonable fit. By contrast, the unshaded model has a wider

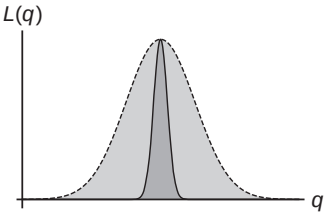

FIGURE 7.9. Illustration of likelihood curves for two classes of models. The peak likelihoods are the same, so the best fit models are equally good, but the widths of the likelihood distributions differ.

range of parameter values that yield an acceptable fit. If we do not have prior reasons to favour a particular parameter range, we have to say that the shaded model is preferred because it has a larger good volume in parameter space. To formalize this idea, we assess the overall ability of a model to fit the data by integrating the likelihood curve (weighted by the priors):

$$Z \equiv p(d) = \int \mathcal{L}(d|q)\ p(q)\ dq. \tag{7.26}$$

The integral yields $p(d)$, or the overall probability of getting the data d from this particular class of models. This quantity is also called the Bayesian 'evidence', Z, and it is what we use to compare different models to each other. Specifically, if we have two models M_1 and M_2, we compare them by examining the ratio known as the Bayes factor:

$$f_{12} = \frac{Z(M_1)}{Z(M_2)}\ \frac{P(M_1)}{P(M_2)},$$

where $P(M_1)$ and $P(M_2)$ are priors on the models themselves. If $f_{12} \gg 1$ then model 1 is favoured, while if $f_{12} \ll 1$ then model 2 is preferred. If f_{12} is comparable to unity then we cannot clearly distinguish between the models. (See Jeffreys 1998 and Section 7.5.2 for more about the quantitative interpretation of Bayes factors.)

This approach provides an objective way to compare models even if they have different numbers of parameters. Integrating over the entire parameter space to compute Z automatically accounts for the fact that models with different numbers of parameters have different volumes.

7.4.6 *Marginalizing parameters*

The posterior distribution contains information about all of the model parameters. What do we do if we want to focus on a subset of the parameters? Formally, we want to obtain the marginal distribution for the parameters of interest by integrating over all the parameters we do not care about (the 'nuisance' parameters).

It might be tempting to avoid integrating by letting the nuisance parameters take on whatever values maximize the posterior, i.e. to optimize rather than marginalize. In general, though, the two steps are not equivalent. Consider a 2D joint probability distribution with the form

$$p(x, y) \propto \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right] \exp\left[-\frac{y^2}{2\sigma_y^2}\right],$$

where $\sigma_y = 1 + x^2$. Suppose y is the nuisance parameter. With optimization, at each value of x we pick the value of y that yields the highest probability (which turns out to

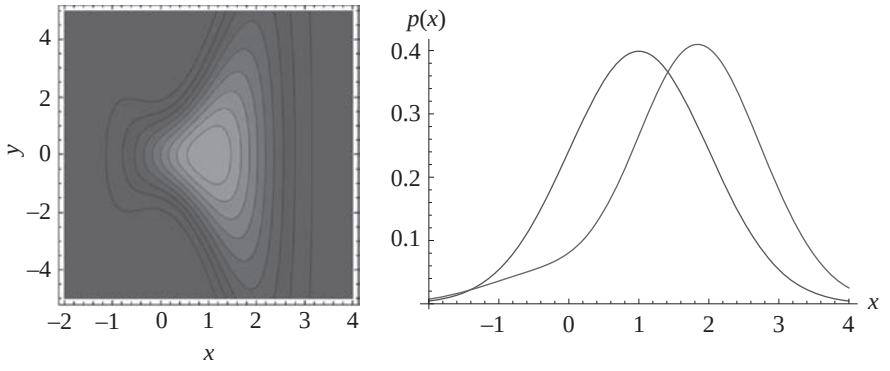

FIGURE 7.10. The left panel shows a sample 2D joint probability distribution $p(x, y)$. In the right panel, the left curve shows the 1D probability distribution $p(x)$ if we merely optimize y, while the right curve shows $p(x)$ if we properly marginalize y.

be $y = 0$ for all x); this would yield the following distribution for x:

$$\text{optimize:} \quad p(x) \propto \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right].$$

With marginalization, we integrate over y and obtain

$$\text{marginalize:} \quad p(x) \propto (1+x^2)\exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right].$$

The inferred distributions for x are clearly different, as shown in Figure 7.10. As x increases, the joint distribution spreads vertically and there is a larger volume in the y-direction that yields a reasonable fit. If we do not have any grounds to prefer particular y values, the availability of a larger volume should lead us to give more weight to larger x values. Marginalizing captures this effect, but optimizing does not. Strictly speaking, we can get away with optimizing only if the posterior is quite narrow in the direction of the nuisance parameters.[†]

7.4.7 *Monte Carlo Markov chains*

All of the preceding discussion is nice in principle, but in practice it may be difficult or even impossible to analyse the full posterior distribution. We can still make progress by using statistical sampling to generate a set of points $\{q_k\}$ drawn from the posterior $p(q|d)$. The popular sampling technique known as Monte Carlo Markov chains (MCMC) can be summarized as follows (e.g. Gelman et al. 2003)[‡]:

- Pick some starting point q_1.
- Postulate some 'trial distribution' $p_{\rm try}(q)$.
- Draw a trial point, $q_{\rm try}$, from $p_{\rm try}(q)$; the probability to accept the trial point is

$$P_{\rm accept} = \min\left[\frac{\mathcal{L}(q_{\rm try})}{\mathcal{L}(q_1)}, 1\right]. \tag{7.27}$$

[†] A posterior that is sufficiently narrow in one direction effectively acts like a Dirac δ-function, causing the integral to pick out the peak value of the parameter.

[‡] To simplify the discussion and notation, we assume flat priors so the posterior is $p(q|d) \propto \mathcal{L}(q)$. Priors can be incorporated by including a factor of $p(q)$ alongside every $\mathcal{L}(q)$.

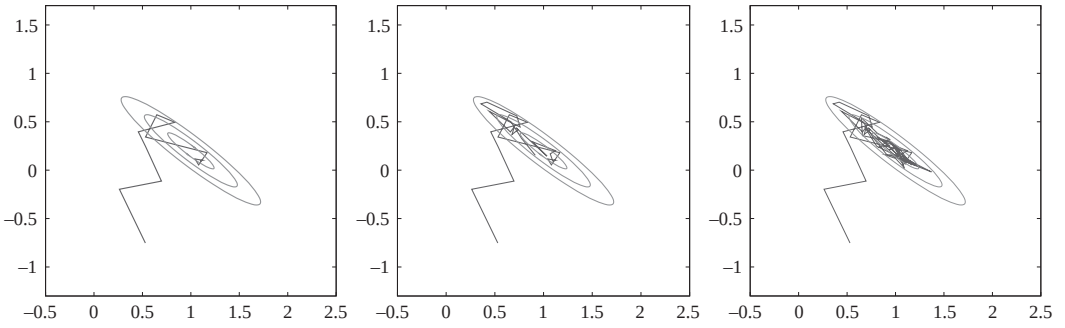

FIGURE 7.11. Illustration of MCMC after 50 (left), 150 (middle) and 250 (right) steps. The oval curves show contours of the likelihood function for this sample case.

- If we accept the trial point, put $q_2 = q_{\rm try}$ (i.e. take a step to $q_{\rm try}$); otherwise, put $q_2 = q_1$ (i.e. stay put).
- Iterate.

The idea behind equation (7.27) is that we definitely want to take steps that increase the likelihood, but we occasionally need to take steps that decrease the likelihood in order to sample the full posterior. Setting the acceptance probability equal to the likelihood ratio (an approach known as the Metropolis–Hastings algorithm; see Metropolis et al. 1953, Hastings 1970) ensures that a large number of iterations will produce a sample of points drawn from $\mathcal{L}(q)$.

To illustrate MCMC, let us consider a sample problem in 2D in which the likelihood is an elongated Gaussian distribution. For now, let the trial distribution be a simple symmetric Gaussian. Figure 7.11 shows the chain of points after 50, 150 and 250 steps. Note that the apparent number of links in the chain is not the same as the number of steps, because sometimes we stay put rather than moving to a new point. (This is particularly common early in the process.) Overall, the chain generally moves towards the peak in the likelihood surface and then moves around exploring the region around the peak.

How do we decide when to stop? A valuable approach is to run multiple independent chains and keep going until the statistical properties of the chains are equivalent. (Gelman et al. 2003 discuss how to evaluate and compare the chains.) Then we throw away the first half of each chain to eliminate any 'memory' of the starting point. Figure 7.12a shows a sample run with ten chains.

How do we pick the trial distribution? In principle, the choice is arbitrary (the key to accurate sampling is the acceptance probability equation (7.27), not the trial distribution). In practice, we should pick a distribution that is easy to work with and reasonably efficient. A Gaussian distribution can be a good first choice, because it yields a sensible mixture of small and large steps, and there are fast algorithms to generate the many Gaussian draws needed for MCMC (e.g. Press et al. 1992). The examples we have seen so far use a Gaussian that is symmetric in x and y. We might be able to do better if we can determine the shape of the likelihood distribution and take bigger steps in directions where $\mathcal{L}(q)$ is elongated while keeping shorter steps in directions where $\mathcal{L}(q)$ is narrow. As the chains advance, we can pause from time to time to compute the covariance matrix of existing points to estimate the shape of the likelihood; we can then use $\mathbf{C}$ to define a multivariate Gaussian trial distribution for the next round of samples. Using 'adaptive'

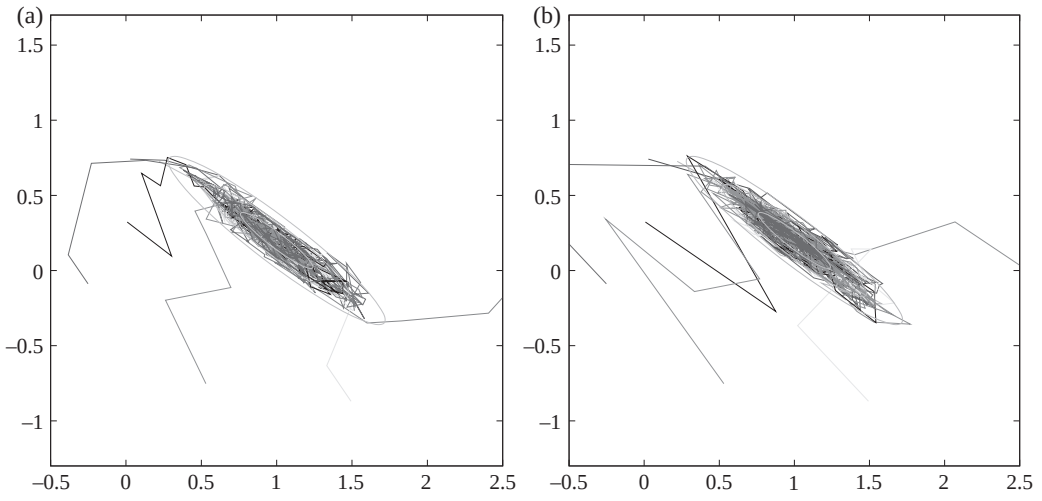

FIGURE 7.12. Illustration of MCMC with 10 independent chains (differentiated in colour; see illustration on website). (a) Simple Gaussian steps. (b) Adaptive steps: the trial distribution is a bivariate Gaussian in which the covariance matrix is updated from time to time using the previously sampled points.

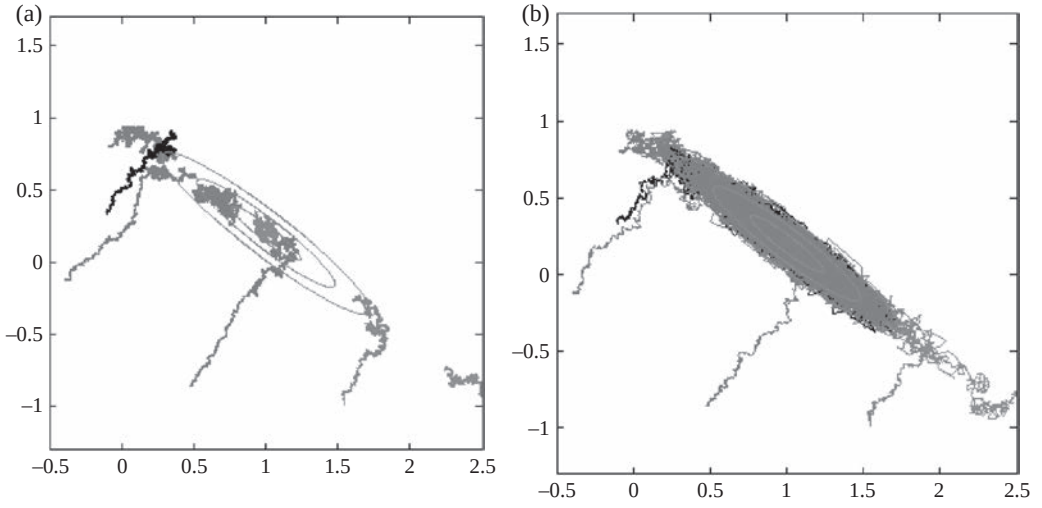

FIGURE 7.13. In (a), the MCMC steps are always small. In (b), the step size is adjusted based on the acceptance rate of trial points.

steps in this way can let MCMC take bigger steps in directions where the likelihood distribution is more elongated and thus explore the volume more quickly (see Figure 7.12b).

That addresses the *shape* of the steps; what about the size? Big steps will often overshoot the likelihood peak, causing MCMC to run for a long time without finding many good points. Small steps have a different problem: it will take an inordinate amount of time to get anywhere, as shown in Figure 7.13a. Fortunately, both of these problems are easy to recognize and remedy. If the steps are too big, the rate at which trial points are accepted will be low (because, again, most steps will overshoot). By contrast, if the steps are too small, the change in likelihood from one step to the next will be small and most trial points will be accepted. Thus, a straightforward solution is to monitor the

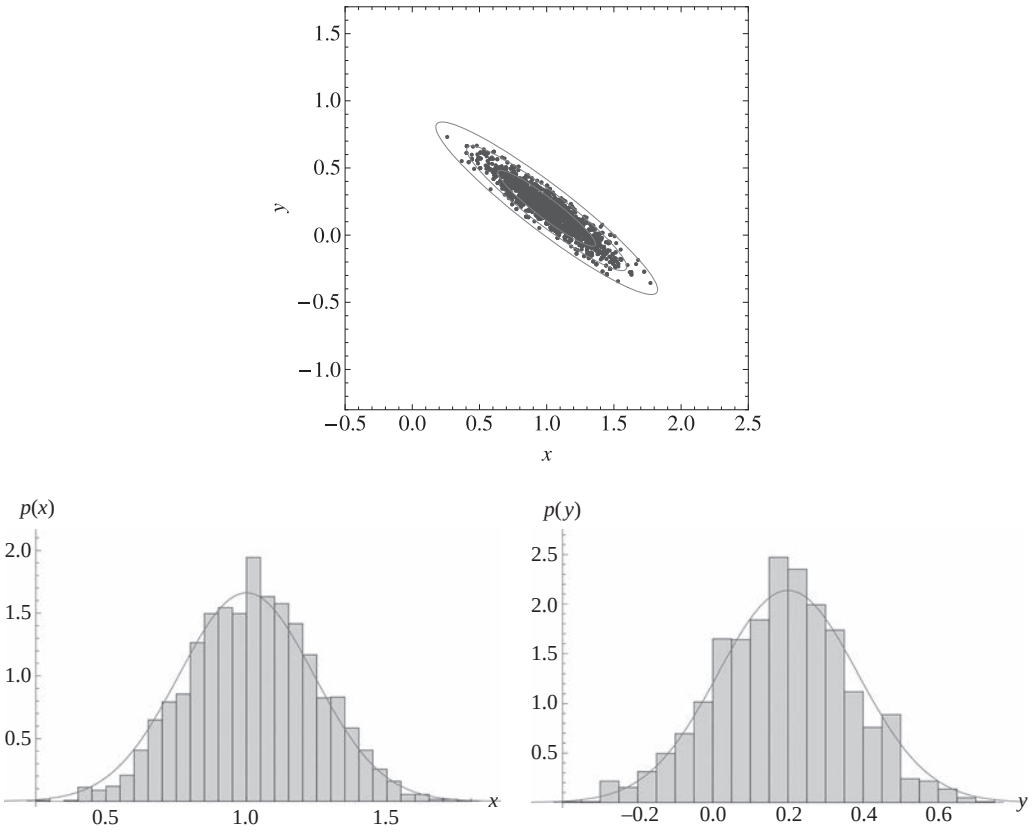

FIGURE 7.14. The top panel shows the joint distribution of sampled points, while the bottom panels show marginal distributions for x and y separately. Ovals indicate the 'truth', while dots indicates the results from MCMC.

acceptance rate: if it is too high, the step size should be increased; if the acceptance rate is too low, the step size should be decreased. Monitoring and adjusting the step size in this way (see Figure 7.13b) can help MCMC automatically find the step size that makes the parameter exploration efficient.

Once we run the chains, what do we do with the results? We can visualize the joint posterior $p(x, y)$ by plotting all of the sampled points in the (x, y) plane (see the top panel of Figure 7.14). We can then see the marginal distribution $p(x)$ by plotting a histogram of all the x values (regardless of the corresponding y values). Similarly, we can obtain $p(y)$ from a histogram of y. In a case with more than two parameters, we can obtain marginal distributions for any subset of the parameters just by plotting a histogram of those parameters.

We can also deal with functions of the model parameters. Suppose we have some function $f(q)$ of the parameters and we want to know its value averaged over the posterior. Formally, we write the average as an integral, but once we have points $\{q_k\}$ drawn from $p(q|d)$ we can turn the integral into a sum:

$$\langle f \rangle = \int f(q)\ p(q|d)\ dq \approx \frac{1}{N} \sum_{k=1}^{N} f(q_k).$$

This is the technique of Monte Carlo integration.

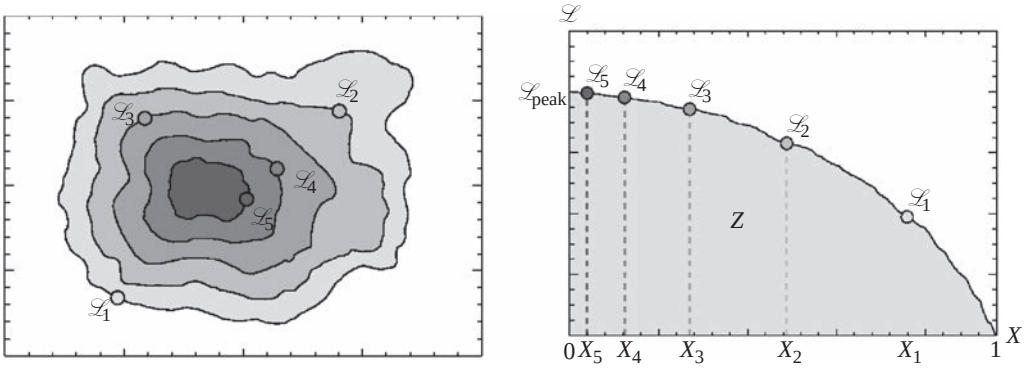

FIGURE 7.15. Schematic depiction of nested sampling. The left panel shows hypothetical likelihood contours. The right panel shows the likelihood as a function of the (fractional) volume enclosed by each contour. The evidence is the area under the $L(X)$ curve. (Courtesy R. Fadely.)

The bottom line is that the points sampled by MCMC capture almost all of the information in the posterior; all we need to do is process the points in different ways.

7.4.8 *Nested sampling*

There is one interesting quantity that is difficult to extract from MCMC: the Bayesian evidence. Since MCMC points are drawn from the posterior, and the total number of points is arbitrary (it depends on how long we run the chains), there is no obvious way to use the chains to evaluate the evidence integral. As an alternative, Skilling (2004, 2006) introduced a technique known as 'nested sampling'. The basic idea, illustrated in Figure 7.15, is to peel away layers of constant likelihood one by one and estimate the volume of each layer statistically. The likelihoods and volumes can be combined to estimate the evidence.

Formally, given a likelihood $\mathcal{L}(q)$ and prior $p(q)$, we can write the evidence integral as[†]

$$Z = \int \mathcal{L}(q)\ p(q)\ dq.$$

Let us define the fractional volume with likelihood higher than L:

$$X(L) = \int_{\mathcal{L}(q)>L} p(q)\ dq.$$

(Note the weighting by the prior.) This is a monotonic decreasing function with $X(0) = 1$ and $X(L_{\rm peak}) = 0$, so we can invert the relation to find $L(X)$ and then write

$$Z = \int_0^1 L(X)\ dX.$$

Now we discretize the integral: if we can find a set of points (L_k, X_k) then we can write

$$Z = \sum_k L_k\,(X_{k-1} - X_k)\,. \tag{7.28}$$

The key to nested sampling is generating the points. Getting the likelihood points is 'easy' (at least conceptually): at step k, we just draw uniformly from the prior in

[†] This formal overview follows Keeton (2011).

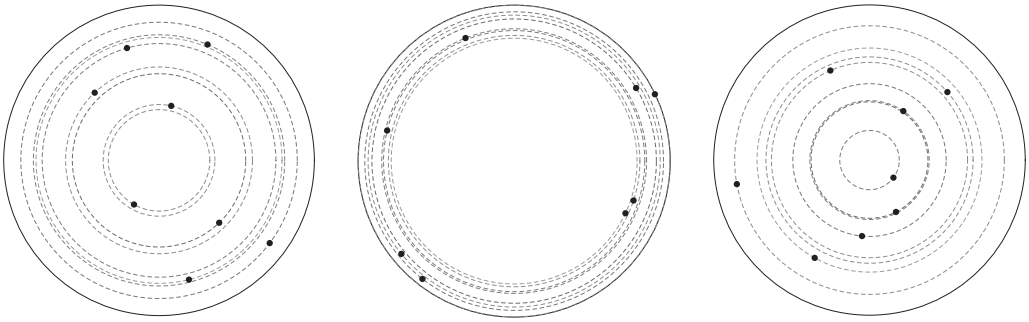

FIGURE 7.16. Three realizations of eight random points inside a circle. The coloured circles indicate likelihood contours. The volumes enclosed by the likelikehood contours are clearly random variables themselves.

the region where $\mathcal{L} > L_{k-1}$. What is harder is getting the associated volumes. In lieu of integration (which we are trying to avoid), we can proceed statistically as follows. Consider a likelihood contour at value L_0 that encloses volume V_0. Now consider M points drawn uniformly within that contour. Let the likelihood contours through those M points enclose volumes $V_1 > V_2 > \cdots > V_M$. We do not know what those volumes are, but we can say they are random variables because the M reference points are themselves random. (Figure 7.16 shows different random realizations.) If we normalize by writing $V_\mu = V_0 t_\mu$ then we know that the $\{t_\mu\}$ are uniform random variables in the range $(0, 1)$. The quantity we most want to know is t_1, which is *the largest of M random numbers drawn uniformly between 0 and 1*. This is a well-known problem in statistics and it is characterized by the probability distribution (e.g. Rose and Smith 2002)

$$p(t) = Mt^{M-1}. \tag{7.29}$$

The mean is $\langle t_1 \rangle = M/(M+1)$.

That reasoning leads to the following algorithm. We begin with M points (known as 'live' points) drawn uniformly from the full prior; let their likelihoods be $\mathcal{L}_\mu$ ($\mu = 1, \ldots, M$). At step k, we extract the live point with the *lowest* value of $\mathcal{L}_\mu$ and call it k-th sampled point:

$$L_k = \min_\mu(\mathcal{L}_\mu).$$

We estimate the associated volume as

$$X_k = X_{k-1} t_k,$$

where t_k is a random number drawn from equation (7.29). We then replace the live point that was extracted by drawing a new point from the priors, restricted to the region with $\mathcal{L}(q) \geq L_k$. We iterate for $N_{\rm nest}$ steps to obtain a set of points (L_k, X_k) that we can use in equation (7.28) to estimate the evidence. (Note: what I have just described is the classic algorithm for nested sampling. Shaw, Bridges and Hobson 2007, Feroz and Hobson 2008, Feroz, Hobson and Bridges 2009, Brewer, Pártay and Csányi Brewer 2010 and Betancourt 2011 have introduced variants that involve different ways of generating the sample points.)

Figure 7.17 shows nested sampling applied to the same test case we saw in Section 7.4.7. The upper left panel shows the starting conditions with 300 live points distributed randomly. The other panels show the situation after 100, 500 or 900 steps. At each step we peel off a layer and estimate its volume in order to build up the evidence as shown in

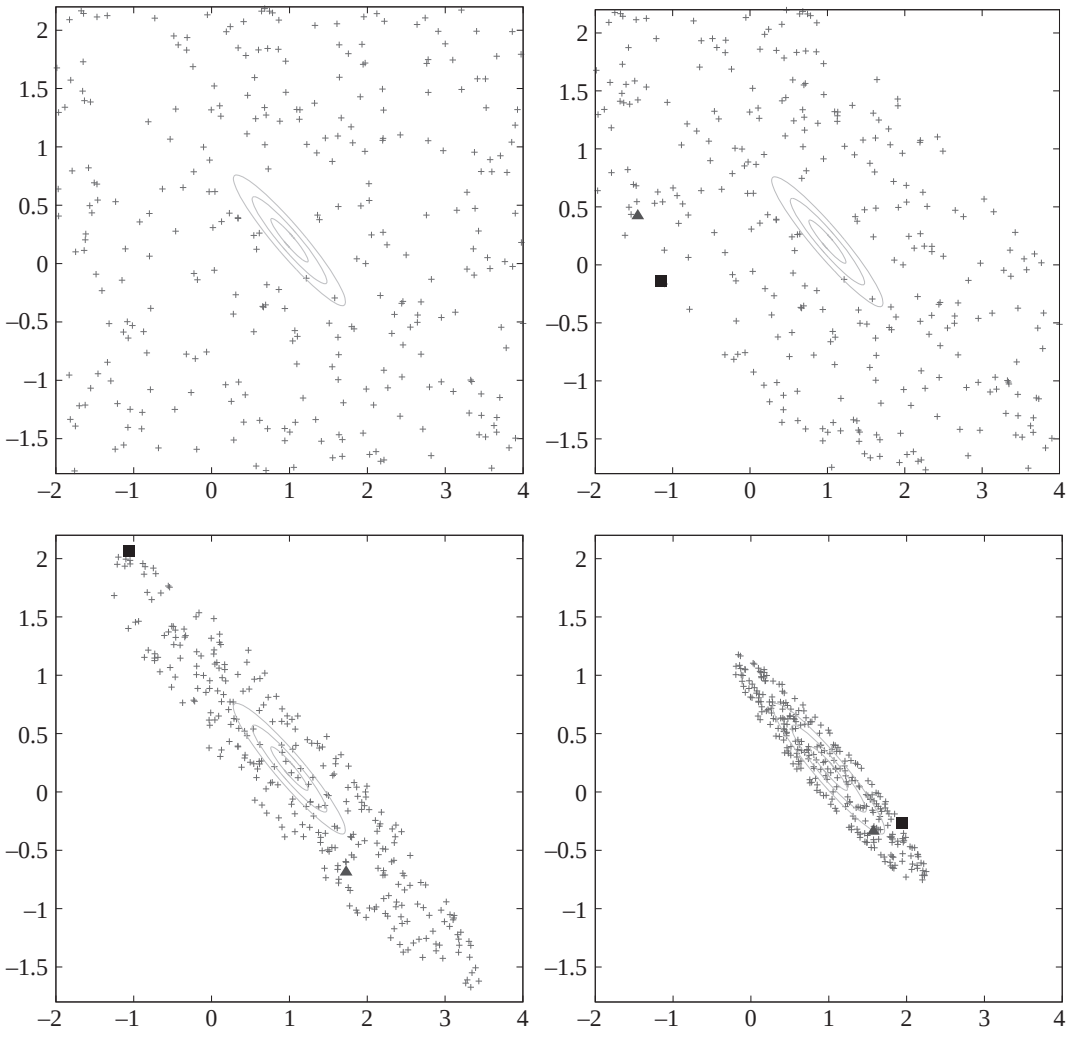

FIGURE 7.17. Illustration of nested sampling. The upper left panel shows the initial set of 300 live points. The other panels show the situation after 100 (upper right), 500 (lower left), or 900 (lower right) steps. At each step, we peel off the outermost live point (black square) and replace it with a new live point (grey triangle) drawn uniformly from the enclosed region.

Figure 7.18. In addition to the sum (7.28), we can estimate a contribution $Z_{\rm live}$ from the live points (basically the average of the $\mathcal{L}_\mu$ values times the volume spanned by the live points). At early stages the net evidence is dominated by $Z_{\rm live}$ and quite noisy because the live points are so spread out. At late stages the evidence becomes robust and $Z_{\rm live}$ goes to zero because the volume spanned by the live points vanishes (while the likelihoods themselves remain finite).

Because it involves random draws, nested sampling is subject to some statistical uncertainty. Keeton (2011) discusses how to compute the statistical uncertainty in the evidence from nested sampling. The amount of uncertainty depends mainly on the number of live points.

The nested sampling process naturally produces a set of points drawn from the parameter space. These points are *not* drawn uniformly from the posterior, so they are not quite

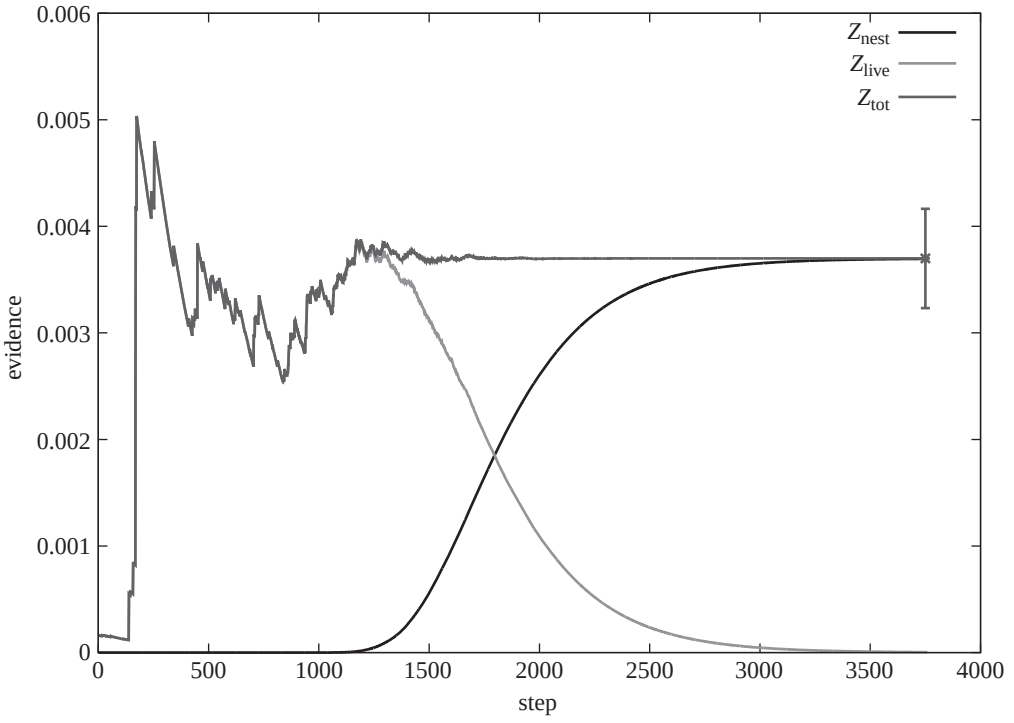

FIGURE 7.18. Build-up of the Bayesian evidence in nested sampling. The horizontal axis is the step number. The black curve shows the contribution from the sum of the sampled points (equation 7.28), while the light grey curve shows the contribution from the live points, and the dark grey curve shows the total. The error bar at the end shows the statistical uncertainty in the evidence computed with the method from Keeton (2011).

as straightforward to interpret as points from MCMC. Nevertheless, the nested sampling points have a known weighting so they can still be used to explore the posterior and do things such as parameter inference (see Skilling 2004, 2006).

7.5. Advanced techniques

Many recent applications of strong lens modelling have gone beyond the basic techniques, trying to capture some of the complexity of real lens mass distributions and/or exploit the information contained in extended images, while giving careful consideration to statistical uncertainties. This section uses two case studies to illustrate advanced versions of the techniques we have already seen, and then gives an overview of additional techniques including free-form mass models, extended sources and line-of-sight effects. This discussion cannot be comprehensive; it is meant to highlight the richness that strong lens modelling can achieve when it is pursued creatively but embedded in a rigorous statistical framework.

7.5.1 *Composite models and astrophysical priors*

The analysis of Q0957+561 by Fadely et al. (2010) illustrates how to build a complex composite model, apply priors from other realms of astrophysics and use MCMC to explore the parameter space. The images in Q0957 are created by the brightest galaxy in a modest-sized cluster of galaxies. The lens galaxy exhibits an ellipticity gradient and

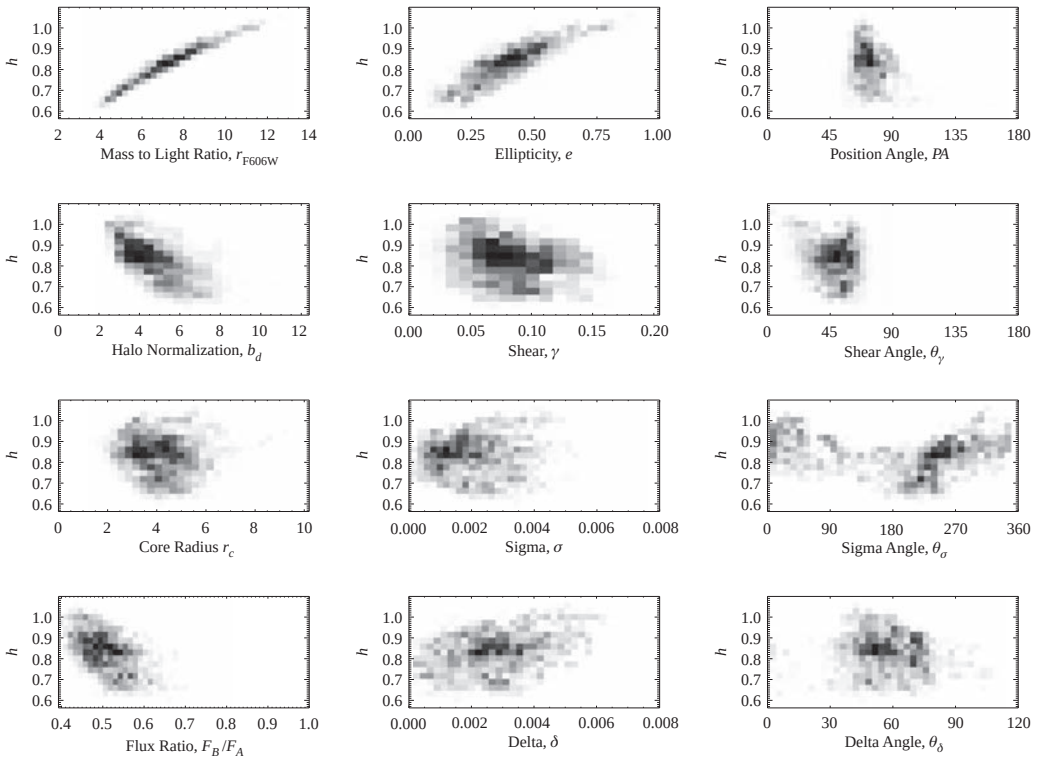

FIGURE 7.19. Two-dimensional histograms depicting the marginalized joint probability distribution $p(q, h)$ for each model parameter q and the scaled Hubble constant h. In the models shown here, the dark matter halo has a softened isothermal profile. (Credit: Fadely et al. 2010; © AAS. Reproduced with permission.)

isophote twist that affect the lens potential on a scale larger than the measurement uncertainties in the lens data (Bernstein and Fischer 1999; Keeton et al. 2000). To incorporate this structure, Fadely et al. included as one model component the observed light distribution scaled by an unknown stellar mass-to-light ratio, Υ. They added a separate dark matter halo modelled as an ellipsoidal mass distribution with a power law or Navarro–Frenk–White (NFW) density profile. The surrounding cluster makes the environment more complicated than usual, so Fadely et al. used a third-order Taylor series expansion:

$$\phi_{\rm env} = \frac{\kappa_c}{2}\, r^2 + \frac{\gamma}{2}\, r^2 \cos 2(\theta - \theta_\gamma) + \frac{\sigma}{4}\, r^3 \cos(\theta - \theta_\sigma) + \frac{\delta}{6}\, r^3 \cos 3(\theta - \theta_\delta). \qquad (7.30)$$

The shear and third-order parameters were fit directly; the mass sheet term, κ_c, was constrained by a separate weak lensing analysis (Nakajima et al. 2009).

To obtain constraints, Fadely et al. examined deep *Hubble Space Telescope* images and catalogued 30 images of 14 distinct sources (some of which had been identified before, and some of which were new). They also included a constraint from the measured time delay between the two quasar images. All told, the set of free parameters included 11 mass model parameters, the Hubble constant, and 28 coordinates for the 14 sources. The set of constraints included 60 coordinates for the 30 images, and the time delay. Fadely et al. optimized the source positions analytically using equation (7.9) and searched the space of 11 mass model parameters using MCMC.

Figure 7.19 shows marginal distributions for the different model parameters in conjunction with the scaled Hubble constant $h = H_0/(100\,{\rm km\,s^{-1}\,Mpc^{-1}})$, assuming a dark

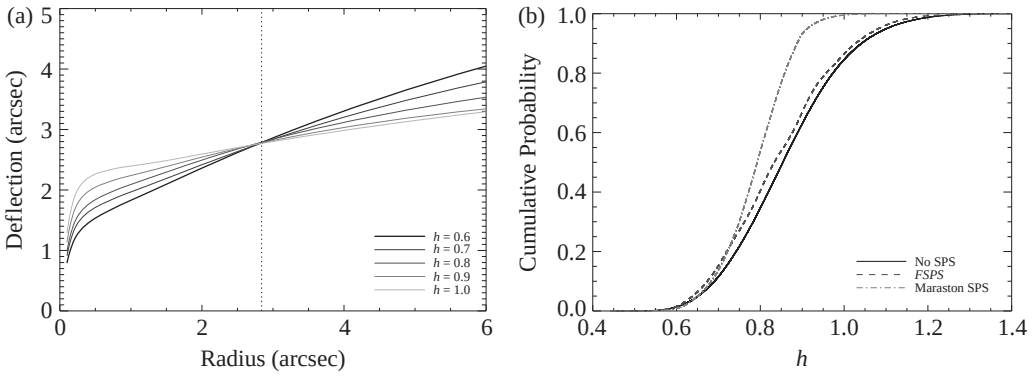

FIGURE 7.20. (a) Deflection curve, $\alpha(r) \propto M(r)/r$, for models with fixed values of h. The dotted line indicates the Einstein radius. (b) Cumulative posterior probability distribution for h with and without constraints on the stellar mass-to-light ratio from stellar population synthesis (SPS) models by Conroy, Gunn and White (2009, 'FSPS') and Maraston et al. (2009). (Credit: Fadely et al. 2010. © AAS. Reproduced with permission.)

matter halo with a softened isothermal profile. (Fadely et al. 2010 show similar plots for other halo models.) Even though there are many constraints, there is also a lot of freedom in the models, so the allowed range of parameters winds up being fairly large. There are clear covariances among parameters; in particular, the model can shift mass between the stellar and dark matter components as long as the total mass within the Einstein radius remains (nearly) constant. Putting more mass into the stellar component leads to a steeper net density profile, which in turn leads to a higher inferred value for the Hubble constant (see Figure 7.20a). Because of this covariance, lensing alone cannot place strong constraints on the density profile and H_0.

The stellar mass-to-light ratio should probably not be treated as completely free, because it can be predicted from stellar population synthesis (SPS) models (e.g. Bruzual and Charlot 2003; Conroy et al. 2009; Maraston et al. 2009). Briefly, the idea of SPS is to use models of stellar evolution and stellar atmospheres to predict how a galaxy spectrum evolves with time. Galaxies that have old stellar populations fade and become redder with age, so fitting SPS models to observed galaxy colours can constrain the stellar mass-to-light ratio. Those constraints can be added as priors in the lens modelling, reducing the final range for h (see Figure 7.20b). From the joint analysis of lensing and SPS models in Q0957, Fadely et al. found $H_0 = 79.3^{+6.7}_{-8.5}$ km s^{-1} Mpc^{-1} at 68% confidence.

In this application, composite parametric models created a flexible framework that revealed the physical context for an important systematic uncertainty in lens models (the radial profile degeneracy) and offered a clear way to use other astrophysical knowledge to break the degeneracy.

7.5.2 *Parametric substructure models*

Composite parametric models can be used in a different way to model dark matter substructure.† The analysis of HE 0435−1223 by Fadely and Keeton (2012) illustrates statistical model comparison based on nested sampling. HE0435 is a four-image lens in

† Vegetti et al. (2010, 2012) use a different approach to constrain dark matter substructure, based on pixelated potential corrections (see Section 7.5.3) applied to lenses with extended sources (see Section 7.5.4).

which the positions and fluxes of the lensed images along with the positions of the main lens galaxy and a nearby neighbour provide 16 constraints. A reasonable smooth lens model has 17 free parameters: the mass, position, ellipticity, position angle, core radius and power law index for the density profile of the main lens galaxy; the mass, position, ellipticity and position angle for the neighbour galaxy; a tidal shear and orientation angle for the rest of the environmental contribution; and the position and flux of the source. The best fit model has $\chi^2 = 24.6$, which is not a good fit – especially when the number of degrees of freedom is negative! Despite being under-constrained, the model apparently lacks some key property that is present in the real lens.

Fadely and Keeton showed that the fit is bad because the smooth model cannot reproduce the brightness of image A. They argued that small-scale structure must modify the flux ratios, so they considered models with mass clumps near the images. Using the Bayesian evidence made it possible to compare models with different numbers of clumps, and hence different numbers of parameters. The following table compares the evidence for clump models to that for the smooth model:

Model	$\log_{10}(Z/Z_{\rm smooth})$
Smooth	$\equiv 0$
1 clump near image A	3.83 ± 0.12
2 clumps near images A and D	3.90 ± 0.13
2 clumps near images A and B	4.46 ± 0.12
3 clumps near images A, B and D	4.35 ± 0.13

According to the Jeffreys (1998) scale, the evidence values provide decisive evidence for a mass clump near image A. The marginalized distributions for the mass of this clump yield

$$\log_{10}\left(M^A_{\rm Ein}\right) = 7.65^{+0.87}_{-0.84} \quad \text{and} \quad \log_{10}\left(M^A_{\rm tot}\right) = 9.31^{+0.44}_{-0.42},$$

where $M_{\rm Ein}$ is the mass within the clump's own Einstein radius, while $M_{\rm tot}$ is the total mass within its truncation radius (in units of $M_\odot$). The evidence values increase further when the models contain an additional mass clump near image B, but according to the Jeffreys scale the increase is not large enough to strongly favour this model. It is interesting that the evidence does not change (within the statistical error bars) with the addition of a clump near image D. The Bayesian framework can reveal that certain parameters add little or no value to a model.

It seems unlikely that the lens galaxy contains just one or two clumps that are closely aligned with the quasar images; it seems more likely that the galaxy has a population of clumps and we are detecting ones that most affect the lensing. Fadely and Keeton examined substructure models using a statistical framework with three sets of parameters (see also Dalal and Kochanek 2002):

- $\mathbf{q}$ = smooth model
- $\mathbf{s}$ = substructure *population* (abundance, mass function, etc.)
- $\mathbf{c}$ = individual clumps (position, mass, etc.)

The likelihood depends on the smooth model and clumps, so it has the form $\mathcal{L}(\mathbf{c}, \mathbf{q})$. The clump properties are drawn from a probability distribution that depends on the population parameters, which we can write as $p(\mathbf{c}|\mathbf{s})$. Finally, we might have priors $p(\mathbf{s}, \mathbf{q})$ on the smooth model and substructure population.

The model may contain hundreds or thousands of clumps, so it is woefully under-constrained. But we do not worry too much, because ultimately what we want is the

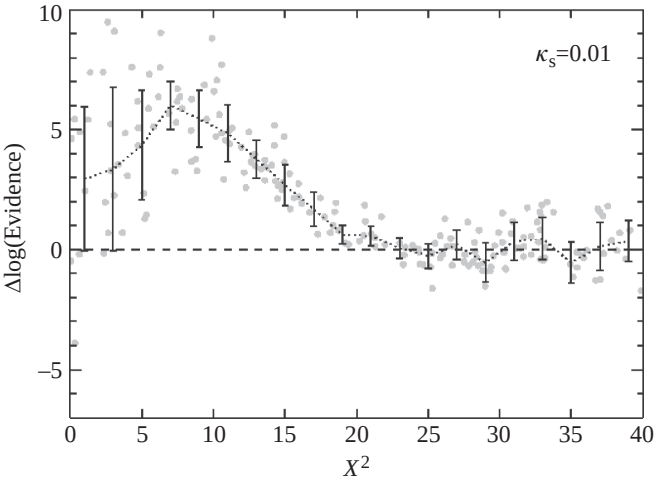

FIGURE 7.21. Each point represents a particular realization of the clump population in models of HE0435. The χ^2 is related to the *peak* likelihood ($\mathcal{L}_{\rm peak} \propto {\rm e}^{-\chi^2/2}$), while the evidence is given by the *integral* over the likelihood. There can be several orders of magnitude scatter in the evidence among models that all have the same χ^2 value, so optimizing is not a good substitute for marginalizing in this case. (Credit: Fadely & Keeton 2012. © 2011 The Authors, *Monthly Notices of the Royal Astronomical Society*, © 2011 RAS.)

marginalized posterior distribution for the substructure population parameters:

$$p(\mathbf{s}) \propto \int \mathcal{L}(\mathbf{c},\mathbf{q})\ p(\mathbf{c}|\mathbf{s})\ p(\mathbf{s},\mathbf{q})\ d\mathbf{c}\ d\mathbf{q}.$$

We have no hope of doing the $\mathbf{c}$ integral explicitly (it may have hundreds or thousands of dimensions), but we can turn to Monte Carlo integration. If $\mathbf{c}_j$ is a realization of the clump population, drawn from $p(\mathbf{c}|\mathbf{s})$, then we can turn the $\mathbf{c}$ integral into a sum[†]:

$$p(\mathbf{s}) \propto \sum_j \int \mathcal{L}(\mathbf{c}_j,\mathbf{q})\ p(\mathbf{s},\mathbf{q})\ d\mathbf{q}.$$

For each $\mathbf{c}_j$, we still need to integrate over $\mathbf{q}$, which can be done with nested sampling. While it might seem easier just to optimize $\mathbf{q}$, recall from Section 7.4.6 that marginalizing and optimizing are not necessarily equivalent. Figure 7.21 shows that they are indeed quite different for HE0435.

Fadely and Keeton assumed the clumps are truncated isothermal spheres that follow a mass function $dN/dm \propto m^{-1.9}$ over the range $m \in 10^7$–$10^{10}\,M_\odot$, and they sought to constrain the mean surface mass density in substructure (Σ_s). Figure 7.22 shows the Bayesian evidence as a function of $\kappa_s = \Sigma_s/\Sigma_{\rm crit}$. Having $\kappa_s \geq 0.001$ increases the evidence by nearly four orders of magnitude relative to models with smaller values (including the smooth model). Translating κ_s values into the substructure mass fraction at the Einstein radius yields the lower limit $f_{\rm sub} > 0.00077$.

This application shows that it is possible to draw valuable conclusions even from models that are under-constrained. There may be parameters that cannot be constrained individually, but if they are not central to the analysis they can be marginalized away.

[†] Dalal and Kochanek (2002) originally introduced this framework for substructure modelling, but they optimized $\mathbf{q}$ rather than marginalizing.

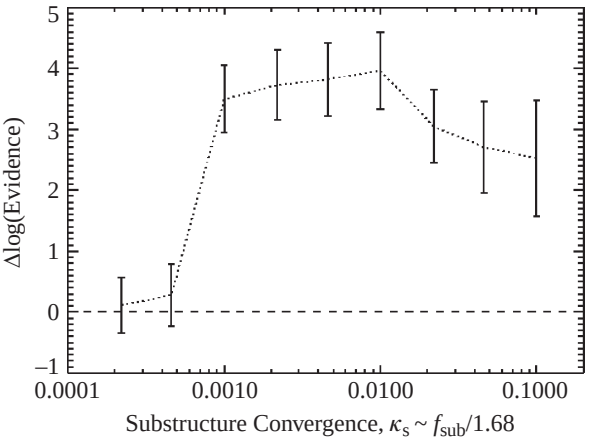

FIGURE 7.22. Differential log evidence (relative to the reference smooth model) as a function of the abundance of substructure, for models of HE0435. (Credit: Fadely & Keeton 2012. © 2011 The Authors, *Monthly Notices of the Royal Astronomical Society*, © 2011 RAS.)

In HE0435, for example, we cannot hope to determine the position and mass of every single clump, but in some sense we do not care (shuffling clumps far from the images has little effect on the model). We can still obtain meaningful constraints on the overall abundance of substructure.

7.5.3 *Free-form mass models*

To this point we have focused on models in which the mass distribution is described by a function intended to resemble real galaxies to some reasonable degree. A different approach is to write the lens potential in terms of some general basis functions,†

$$\phi(\mathbf{x}) = \sum_\nu a_\nu \, f_\nu(\mathbf{x}), \tag{7.31}$$

and then fit for the expansion coefficients. Since the lens equation and time delay are linear in the potential and its derivatives, the image positions and time delays provide a set of constraint equations that are linear in a_ν, $\mathbf{u}$ and t_0^{-1}. (Flux constraints and shear can introduce non-linearities, but being able to handle a significant portion of the problem using linear techniques still helps.)

If the constraints outnumber the unknowns, the models are over-constrained and the analysis is similar to what we have seen for parametric models. If the models are under-constrained, though, the approach changes. For a model with $\nu < 0$ degrees of freedom, the system of constraint equations has a *space* of solutions with dimension $|\nu|$. Much of the solution space will correspond to models that are physically impossible (e.g. they have regions of negative mass) or at least implausible (e.g. they have structures that do not make sense in terms of real astrophysical objects). To eliminate such solutions, and generally narrow the allowed range, we can impose priors on the models. Let us briefly examine different types of free-form models and the priors imposed on them.

Multipole models. Perhaps the simplest type of free-form model assumes a radial profile that corresponds to an isothermal density, but allows a general angular structure

† This section provides a summary of the more extended discussion by Keeton (2010).

by using a multipole expansion:

$$\phi(r,\theta) = r \sum_{m=0}^{m_{\max}} (a_m \cos m\theta + b_m \sin m\theta) .$$

Evans and Witt (2003) chose $m_{\max}$ to make the number of unknowns match the number of constraints, which is a type of prior. Congdon and Keeton (2005) let the models be under-constrained and looked for solutions with the smallest deviations from elliptical symmetry (the least 'wiggles' in isodensity contours). Yoo et al. (2005, 2006) used multipole models to fit Einstein rings, which offered enough constraints to make the models over-constrained.

Multipole/Taylor models. At the next level of complexity, we keep the multipole expansion for the angular structure but allow a more general radial profile. Trotter, Winn and Hewitt (2000) pointed out that the images are often 'near' the Einstein radius, so we can consider a Taylor series expansion in $r - r_0$ (or equivalently $r/r_0 - 1$):

$$\phi(r,\theta) = \sum_{m=0}^{m_{\max}} \sum_{n=0}^{n_{\max}} \left(\frac{r}{r_0} - 1\right)^n (a_{mn} \cos m\theta + b_{mn} \sin m\theta) .$$

Trotter, Winn and Hewitt (2000) applied this approach to MG J0414+0534, which has more than the usual number of constraints because the images can be resolved into multiple subcomponents by high-resolution radio observations (Ros et al. 2000). The extra constraints allowed Trotter et al. to include a reasonable number of terms in the multipole/Taylor expansion and still keep the problem over-constrained.

Pixelated mass maps. Saha and Williams introduced a very flexible approach based on mass pixels (Saha and Williams 1997; Williams and Saha 2000; Saha and Williams 2004). This yields models with hundreds or even thousands of unknowns, so it is very dependent on priors. Saha and Williams impose the following priors: (a) the density must be non-negative; (b) the density gradient must point within 45° of the lens centre; (c) no pixel value may exceed the average of its neighbours by more than a factor of 2 (except for the central pixel); (d) the projected density profile must be steeper than $r^{-1/2}$. Also, if desired, the mass map can be required to have inversion symmetry.

Those priors eliminate models that are grossly unphysical but still allow a range of solutions that may be too broad. For example, the surface mass density could be non-negative yet the corresponding 3D density could go negative in places. Also, the models are not guaranteed to produce the correct number of images. Even with the priors, many models may have shapes that are implausible. Pixelated mass models therefore yield what is probably an overly generous range of allowed models. That said, the models can be used (with some care) to draw valuable conclusions about the physical properties of lens galaxies and the Hubble constant (e.g. Saha et al. 2006; Ferreras, Saha and Williams 2005; Ferreras et al. 2010; Leier et al. 2011).

Pixelated potential corrections. Several groups (Blandford, Surpi and Kundić 2001; Suyu and Blandford 2006; Koopmans 2005; Vegetti and Koopmans 2009) have combined the parametric and free-form approaches by writing the lens potential as $\phi = \phi_0 + \delta\phi$ where ϕ_0 is a parametric model that is designed to capture most of the properties of the lens, while $\delta\phi$ consists of corrections that allow more generality. For a fixed ϕ_0, the correction terms $\delta\phi$ can be evaluated on a grid of pixels using linear techniques; the parametric piece can then be varied and the process iterated to find the best fit. These

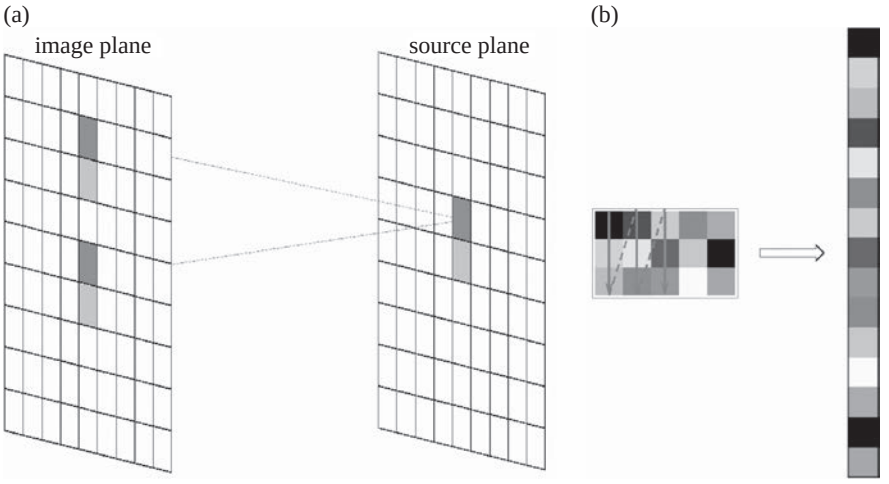

FIGURE 7.23. (a) Since lensing conserves surface brightness, the surface brightness values of pixels in the image plane are equal to the values of the corresponding pixels in the source plane. This is true even if there are multiple images of the source pixels. (b) We can treat a 2D image as a vector simply by stacking columns. Thus, each plane in the panel on the left can be treated as a vector.

models have been applied with particular success to lenses with extended sources (Suyu et al. 2009, 2010; Vegetti et al. 2010, 2012).

7.5.4 *Extended sources*

When lenses have extended images (e.g. Bolton et al. 2008; Brownstein et al. 2012; Stark et al. 2013), there are many constraints from all the pixel brightnesses, but also many free parameters because we do not know the intrinsic structure of the source. Specialized techniques are required to reconstruct the source as part of the modelling process (Wallington et al. 1994, 1995, 1996; Warren and Dye 2003; Treu and Koopmans 2004; Dye and Warren 2005; Koopmans 2005; Suyu et al. 2006; Wayth and Webster 2006; Vegetti and Koopmans 2009; Tagore and Keeton 2014).

General source reconstruction methods rest on the principle that lensing conserves surface brightness.[†] In an idealized setting with no smearing effects from the atmosphere or telescope, the surface brightness values of pixels in the image plane are exactly equal to the values of the corresponding pixels in the source plane (see Figure 7.23a). In other words, there is a linear mapping between the array of pixel values in the source plane and the array of pixel values in the image plane. If we treat each pixel map as a vector (by stacking columns, as in Figure 7.23b), we can write the relation between the source vector $\mathbf{s}$ and the image vector $\mathbf{d}$ as

$$\mathbf{d} = \mathbf{L}_0 \, \mathbf{s},$$

where $\mathbf{L}_0$ is a matrix known as the 'lensing operator'. In principle $\mathbf{L}_0$ could contain just 0's and 1's, although in practice it often contains fractional values to account for interpolation. This framework can be extended to include smearing effects. Since those just redistribute photons, they can be described by a 'blurring' operator, $\mathbf{B}$, that is also

[†] This presentation draws on Keeton (2010).

linear. The net mapping has the form

$$\mathbf{d} = \mathbf{L}\,\mathbf{s}, \tag{7.32}$$

where $\mathbf{L} = \mathbf{B}\,\mathbf{L}_0$.

If we observe an image characterized by data $\mathbf{d}^{\rm obs}$ and covariance matrix $\mathbf{S}_d$, we can define a goodness of fit

$$\chi^2_{\rm img} = \left(\mathbf{L}\,\mathbf{s} - \mathbf{d}^{\rm obs}\right)^t\,\mathbf{S}_d^{-1}\,\left(\mathbf{L}\,\mathbf{s} - \mathbf{d}^{\rm obs}\right).$$

Depending on how we pixelate the source and image planes, there may be (many) more parameters than constraints, and thus a large family of solutions. Many of the recovered sources may be unphysical (e.g. negative flux) or implausible (e.g. spikes or weird shapes), so we need to find a way to eliminate or at least penalize those. This approach is known as 'regularizing' the solution, and in Bayesian language it represents a sort of prior. Suppose we want to penalize a model with spikes. We might add a χ^2 term that gets large when too many pixel values become too large:

$$\chi^2_{\rm reg} \;\sim\; \sum s_j^2 = \mathbf{s}^t\,\mathbf{s}.$$

Alternatively, we might want to penalize large gradients in the surface brightness distribution (e.g. from sharp edges). We can write a vector of first derivatives in the form $\mathbf{v} = \mathbf{H}_v\mathbf{s}$, where $\mathbf{H}_v$ is an operator that computes derivatives using finite differencing. In this case, the penalty term has the form

$$\chi^2_{\rm reg} \;\sim\; \mathbf{v}^t\mathbf{v} \;\sim\; \mathbf{s}^t\,\mathbf{H}_v^t\,\mathbf{H}_v\,\mathbf{s}.$$

Still a third possibility is to penalize strong curvature terms. Computing second derivatives with finite differencing leads to a penalty term of the form

$$\chi^2_{\rm reg} \;\sim\; \mathbf{s}^t\,\mathbf{H}_a^t\,\mathbf{H}_a\,\mathbf{s}.$$

In other words, we can handle various types of regularization by adding a term of the form $\mathbf{s}^t\,\mathbf{H}^t\,\mathbf{H}\,\mathbf{s}$ to the goodness of fit (Warren and Dye 2003; Treu and Koopmans 2004; Dye and Warren 2005; Koopmans 2005; Suyu et al. 2006; Vegetti and Koopmans 2009; Tagore and Keeton 2014):

$$\chi^2 = \left(\mathbf{L}\,\mathbf{s} - \mathbf{d}^{\rm obs}\right)^t\,\mathbf{S}_d^{-1}\,\left(\mathbf{L}\,\mathbf{s} - \mathbf{d}^{\rm obs}\right) + \lambda_s\,\mathbf{s}^t\,\mathbf{H}^t\,\mathbf{H}\,\mathbf{s}, \tag{7.33}$$

where the structure of the $\mathbf{H}$ matrix determines the type of regularization. Also, λ_s is the 'regularization strength' such that a low λ_s puts more emphasis on obtaining a good fit while a high λ_s puts more emphasis on having a reasonable source.

For a given value of λ_s, the optimal source can be found by solving $\nabla_{\mathbf{s}}\chi^2 = 0$, or

$$\left(\mathbf{L}^t\,\mathbf{S}_d^{-1}\,\mathbf{L} + \lambda_s\,\mathbf{H}^t\,\mathbf{H}\right)\,\mathbf{s} = \mathbf{L}^t\,\mathbf{S}_d^{-1}\,\mathbf{d}^{\rm obs}.$$

To finish the analysis we need to set the regularization strength. One possibility is to decide in advance how strongly to regularize and apply an appropriate prior on λ_s. Another possibility is to treat λ_s as a nuisance parameter and marginalize it. Suyu et al. (2006) argue that the posterior has a fairly sharp peak as a function of λ_s, so simply optimizing may be adequate.

The ability to fit lenses with extended sources has enabled a vast array of applications relating to the physical properties of lens galaxies (e.g. Koopmans et al. 2006), dark

matter substructure (e.g. Vegetti et al. 2010, 2012), and cosmological parameters (Suyu et al. 2010, 2013).

7.5.5 *Line-of-sight effects*

We end with a few words about extending lens modelling into the third dimension along the line of sight (LOS). In most strong lens systems the light bending is dominated by mass in a single lens plane, but it may be perturbed by other objects as the light travels through the Universe. We need to generalize the theoretical framework to handle multiple deflections. If there is mass in N planes (labelled in order of increasing redshift), the multiplane lens equation has a recursive form such that the position of a light ray in plane j is (Blandford and Narayan 1986; Kovner 1987; Schneider, Ehlers and Falco 1992; Petters, Levine and Wambsganss 2001)

$$\mathbf{x}_j = \mathbf{x}_1 - \sum_{i=1}^{j-1} \beta_{ij} \alpha_i \left(\mathbf{x}_i \right) , \tag{7.34}$$

where α_i is the deflection caused by mass in plane i, and β_{ij} is a ratio of angular diameter distances between various planes:

$$\beta_{ij} = \frac{D_{ij} D_s}{D_j D_{is}} .$$

The image position on the sky is $\mathbf{x}_1$, and the source position is $\mathbf{x}_s = \mathbf{x}_{N+1}$. Suppose there is one main lens galaxy with index ℓ, and all of the other galaxies lie sufficiently far from the lens (in projection) that we can expand their contributions in Taylor series of the form equation (7.17). Then we can collect all of the perturbations into a set of matrices and write the lens equation as (Kovner 1987; Schneider et al. 1992)

$$\mathbf{x}_s = \mathbf{B}_s \mathbf{x}_1 - \mathbf{C}_{\ell s} \boldsymbol{\alpha}_\ell \left(\mathbf{B}_\ell \mathbf{x}_1 \right) , \tag{7.35}$$

where $\mathbf{B}_\ell$ is a weighted sum of contributions from perturbers in the foreground, $\mathbf{C}_{\ell s}$ is a weighted sum of contributions from perturbers in the background, and $\mathbf{B}_s$ is a (different) weighted sum of contributions from all perturbers. If some of the perturbers are close enough to the lens that the shear approximation is not sufficiently accurate, we can develop a hybrid of equations (7.34) and (7.35) that treats some perturbers exactly and others with the multiplane shear approximation (McCully et al. 2014).

There is growing evidence that detailed studies of gravitational lenses need to account for LOS effects. One line of evidence comes from using pencil-beam redshift surveys to map galaxies and build models of the 3D mass distributions along the lines of sight to real lenses (Momcheva et al. 2006; Williams et al. 2006; Wong et al. 2011). LOS contributions to the shear are non-negligible, and they are different from a shear in the main lens plane. Lens models that assume all the shear lies in the same plane can have both bias and scatter in the recovered model parameters (McCully et al., in preparation). A second line of evidence comes from ray tracing through numerical simulations along trajectories that resemble lens fields (Hilbert et al. 2009; Collett et al. 2013; Greene et al. 2013). Such analyses can be used to calibrate the probability distribution for LOS effects, which remain one of the most important systematic uncertainties in attempts to constrain cosmological parameters with lensing (e.g. Suyu et al. 2013). The bottom line is that lens modelling needs to shift away from the traditional 2D view and begin to account for 3D effects from the line of sight as the data and modelling methods continue to improve.

Acknowledgements

I thank the organizers of the XXIV Winter School for the opportunity to participate. I thank Ross Fadely, Curtis McCully, Amit Tagore, Ken Wong and Ann Zabludoff for their collaboration on aspects of the methodology discussed here, and Lisa Fishenfeld for comments on the manuscript. This work received support from the US National Science Foundation through grants AST-0747311 and AST-1211385.

REFERENCES

Bernstein, G. & Fischer, P. 1999, *AJ*, **118**, 14

Betancourt, M. 2011, in AIP Conf. Ser. 1305, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Proc. 30th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, ed. A. Mohammad-Djafari, J.-F. Bercher & P. Bessière (New York: AIP), 165

Blandford, R. & Narayan, R. 1986, *ApJ*, **310**, 568

Blandford, R., Surpi, G. & Kundić, T. 2001, in ASP Conf. Ser. 237, *Gravitational Lensing: Recent Progress and Future Goals*, ed. T. G. Brainerd & C. S. Kochanek (San Francisco: ASP), 65

Bolton, A. S., Burles, S., Koopmans, L. V. E., Treu, T., Gavazzi, R., Moustakas, L. A., Wayth, R. & Schlegel, D. J. 2008, *ApJ*, **682**, 964

Brewer, B. J., Pártay, L. B. & Csányi, G. 2010, DNEST: Diffusive Nested Sampling. Astrophysics Source Code Library

Brownstein, J. R. et al. 2012, *ApJ*, **744**, 41

Bruzual, G. & Charflot, S. 2003, *MNRAS*, **344**, 1000

Collett, T. E. et al. 2013, *MNRAS*, **432**, 679

Congdon, A. B. & Keeton, C. R. 2005, *MNRAS*, **364**, 1459

Conroy, C., Gunn, J. E. & White, M. 2009, *ApJ*, **699**, 486

Dalal, N. & Kochanek, C. S. 2002, *ApJ*, **572**, 25

Dye, S. & Warren, S. J. 2005, *ApJ*, **623**, 31

Evans, N. W. & Witt, H. J. 2003, *MNRAS*, **345**, 1351

Fadely, R. & Keeton, C. R. 2012, *MNRAS*, **419**, 936

Fadely, R., Keeton, C. R., Nakajima, R. & Bernstein, G. M. 2010, *ApJ*, **711**, 246

Falco, E. E., Gorenstein, M. V. & Shapiro, I. I. 1985, *ApJL*, **289**, L1

Feroz, F. and Hobson, M. P. 2008, *MNRAS*, **384**, 449

Feroz, F., Hobson, M. P. & Bridges, M. 2009, *MNRAS*, **398**,1601

Ferreras, I., Saha, P. & Williams, L. L. R. 2005, *ApJL*, **623**, L5

Ferreras, I., Saha, P., Leier, D., Courbin, F. & Falco, E. E. 2010, *MNRAS*, **409**, L30

Gelman, A., Carlin, J. B., Stern, H. & Rubin, D. B. 2003, *Bayesian Data Analysis* (Boca Raton: Chapman & Hall/CRC)

Gorenstein, M. V., Shapiro, I. I. & Falco, E. E. 1988, *ApJ*, **327**, 693

Greene, Z. S. et al. 2013, *ApJ*, **768**, 39

Hastings, W. K. 1970, *Biometrika*, **57(1)**, 97

Hilbert, S., Hartlap, J., White, S. D. M. & Schneider, P. 2009, *A&A*, **499**, 31

Jeffreys, H. 1998, *Theory of Probability*, 3rd edn (Oxford: Oxford University Press)

Keeton, C. R. 2001, A Catalog of Mass Models for Gravitational Lensing, arXiv:astro-ph/0102341

Keeton, C. R. 2010, *General Relativity and Gravitation*, **42**, 2151

Keeton, C. R. 2011, *MNRAS*, **414**, 1418

Keeton, C. R., Falco, E. E., Impey, C. D., Kochanek, C. S., Lehár, J., McLeod, B. A., Rix, H.-W., Muñoz, J. A. & Peng, C. Y. 2000, *ApJ*, **542**, 74

Kochanek, C. S. 1991, *ApJ*, **373**, 354

Koopmans, L. V. E. 2005, *MNRAS*, **363**, 1136

Koopmans, L. V. E., Treu, T., Bolton, A. S., Burles, S. & Moustakas, L. A. 2006, *ApJ*, **649**, 599

Kovner, I. 1987, *ApJ*, **316**, 52

Leier, D., Ferreras, I., Saha, P. & Falco, E. E. 2011, *ApJ*, **740**, 97

McCully, C., Keeton, C. R., Wong, K. C. & Zabludoff, A. I. 2014, *MNRAS*, **443**, 3631

Maraston, C., Strömbäck, G., Thomas, D., Wake, D. A. & Nichol, R. C. 2009, *MNRAS*, **394**, l107

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. 1953, *J. Chem. Phys.*, **21**, 1087

Momcheva, I., Williams, K., Keeton, C. & Zabludoff, A. 2006, *ApJ*, **641**, 169

Nakajima, R., Bernstein, G. M., Fadely, R., Keeton, C. R. & Schrabback, T. 2009, *ApJ*, **697**, 1793

Petters, A. O., Levine, H. & Wambsganss, J. 2001, *Singularity Theory and Gravitational Lensing* (Boston: Birkhäuser)

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. 1992, *The Art of Scientific Computing*, 2nd edn (Cambridge: Cambridge University Press)

Ros, E., Guirado, J. C., Marcaide, J. M., Pérez-Torres, M. A., Falco, E. E., Muñoz, J. A., Alberdi, A. & Lara, L. 2000, *A&A*, **362**, 845

Rose, C. & Smith, M. 2002, *Mathematical Statistics with Mathematica* (Berlin: Springer-Verlag)

Ross, S. 2012, *A First Course in Probability*, 9th edn (New Jersey: Pearson Education)

Saha, P. & Williams, L. L. R. 1997, *MNRAS*, **292**, 148

Saha, P. & Williams, L. L. R. 2004, *AJ*, **127**, 2604

Saha, P., Coles, J., Macciò, A. V. & Williams, L. L. R. 2006, *ApJL*, **650**, L17

Schneider, P., Ehlers, J. & Falco, E. E. 1992, *Gravitational Lenses* (Berlin: Springer-Verlag)

Shaw, J. R., Bridges, M. & Hobson, M. P. 2007, *MNRAS*, **378**, 1365

Shewchuk, J. R. 1996, in Lecture Notes in Computer Science 1148, *Applied Computational Geometry: Towards Geometric Engineering*, ed. M. C. Lin & D. Manocha (Berlin: Springer-Verlag)

Shewchuk, J. R. 2002, *Computational Geometry*, **22(1–3)**, 21

Skilling, J. 2004, in AIP Conf. Ser. 735, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering* (New York: AIP), 395

Skilling, J. 2006, *Bayesian Analysis*, **1**, 833

Stark, D. P. et al. 2013, *MNRAS*, **436**, 1040

Suyu, S. H. & Blandford, R. D. 2006, *MNRAS*, **366**, 39

Suyu, S. H., Marshall, P. J., Hobson, M. P. & Blandford, R. D. 2006, *MNRAS*, **371**, 983

Suyu, S. H., Marshall, P. J., Blandford, R. D., Fassnacht, C. D., Koopmans, L. V. E., McKean, J. P. & Treu, T. 2009, *ApJ*, **691**, 277

Suyu, S. H., Marshall, P. J., Auger, M. W., Hilbert, S., Blandford, R. D., Koopmans, L. V. E., Fassnacht, C. D. & Treu, T. 2010, *ApJ*, **711**, 201

Suyu, S. H. et al. 2013, *ApJ*, **766**, 70

Tagore, A. S. & Keeton, C. R. 2014, *MNRAS*, **445**, 694

Treu, T. & Koopmans, L. V. E. 2004, *ApJ*, **611**, 739

Trotter, C. S., Winn, J. N. & Hewitt, J. N. 2000, *ApJ*, **535**, 671

Vegetti, S. & Koopmans, L. V. E. 2009, *MNRAS*, **392**, 945

Vegetti, S., Koopmans, L. V. E., Bolton, A., Treu, T. & Gavazzi, R. 2010, *MNRAS*, **408**, 1969

Vegetti, S., Lagattuta, D. J., McKean, J. P., Auger, M. W., Fassnacht, C. D. & Koopmans, L. V. E. 2012, *Nature*, **481**, 341

Wallington, S., Narayan, R. & Kochanek, C. S. 1994, *ApJ*, **426**, 60

Wallington, S., Kochanek, C. S. & Koo, D. C. 1995, *ApJ*, **441**, 58

Wallington, S., Kochanek, C. S. & Narayan, R. 1996, *ApJ*, **465**, 64

Warren, S. J. & Dye, S. 2003, *ApJ*, **590**, 673

Wayth, R. B. & Webster, R. L. 2006, *MNRAS*, **372**, 1187

Williams, K. A., Momcheva, I., Keeton, C. R., Zabludoff, A. I. & Lehár, J. 2006, *ApJ*, **646**, 85

Williams, L. L. R. & Saha, P. 2000, *AJ*, **119**, 439

Wong, K. C., Keeton, C. R., Williams, K. A., Momcheva, I. G. & Zabludoff, A. I. 2011, *ApJ*, **726**, 84

Yoo, J., Kochanek, C. S., Falco, E. E. & McLeod, B. A. 2005, *ApJ*, **626**, 51

Yoo, J., Kochanek, C. S., Falco, E. E. & McLeod, B. A. 2006, *ApJ*, **642**, 22